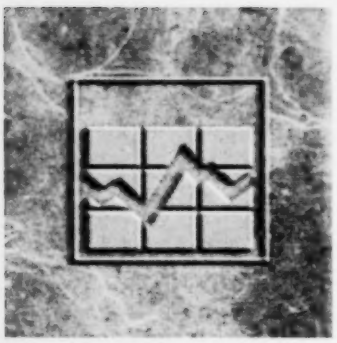
N° 75F0002MIF au catalogue — N° 007
ISSN 1707-2867
ISBN 978-0-662-07427-4

Document de recherche

Série de documents de recherche - Revenu

Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu

2003



par Sylvie LaRoche

Division des méthodes d'enquêtes ménages
Immeuble Jean-Talon, Ottawa, K1A 0T6

Téléphone: 1-613-951-7355

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division de la statistique du revenu, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 613-951-7355; 888-297-7355: revenu@statcan.ca).

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.ca ou par téléphone entre 8h30 et 16h30 du lundi au vendredi aux numéros suivants :

Numéros sans frais (Canada et États-Unis) :	
Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369
Renseignements concernant le Programme des services de dépôt	1-800-635-7943
Télécopieur pour le Programme des services de dépôt	1-800-565-7757
Centre de renseignements de Statistique Canada :	1-613-951-8116
Télécopieur	1-613-951-0581

Renseignements pour accéder au produit

Le produit nº 75F0002MIF au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Publications.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui sont observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.ca sous À propos de nous > Offrir des services aux Canadiens.

2008-02973/

Statistique Canada
Division de la statistique du revenu

Série de documents de recherche - Revenu

Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu

2003

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2007

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Octobre 2007

Nº 75F0002MIF au catalogue, nº 7
ISSN 1707-2867
ISBN 978-0-662-07427-4

Périodicité : hors série

Ottawa

This publication is available in English upon request (catalogue no. 75F0002MIF).

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Résumé

L'Enquête sur la Dynamique du travail et du revenu (EDTR), introduite à l'année de référence 1993, est une enquête longitudinale par panel menée auprès des individus. Elle vise à mesurer les changements au niveau du bien-être économique des individus et les facteurs qui peuvent influencer ces changements. L'échantillon de l'EDTR est divisé en deux panels qui se chevauchent, d'une durée de six ans chacun. Les enquêtes longitudinales, comme l'EDTR, sont particulières de par la nature dynamique de la composition de leur échantillon, engendrée directement par la dynamique des familles et des ménages au fil des ans. À chaque année de référence, l'EDTR produit deux ensembles de poids : un ensemble de poids qui est représentatif de la population initiale d'un panel (l'aspect longitudinal) et un autre qui est représentatif de la population actuelle (l'aspect transversal). Depuis 2002, un troisième ensemble de poids combinant deux panels se chevauchant pour former un nouvel échantillon longitudinal est aussi produit (appelé pondération longitudinale combinée).

Pour la production de poids transversaux, l'EDTR combine deux échantillons indépendants et assigne un poids aux individus qui se sont joints à l'échantillon après la sélection de l'échantillon initial. Les poids longitudinaux tout comme les poids transversaux sont ajustés pour la non-réponse et les valeurs influentes. De plus, un redressement de l'échantillon est effectué pour représenter la population cible. Le but de ce document est donc de décrire la méthodologie employée à l'EDTR pour pondérer les échantillons longitudinal et transversal, et de présenter les développements importants à venir. Afin de mieux illustrer la stratégie de pondération, les résultats de l'année de référence 2003 sont utilisés. Il est à noter que la méthodologie employée pour la pondération longitudinale combinée ne sera pas présentée dans ce document puisqu'elle est décrite de façon exhaustive dans Naud (2004).

TABLE DES MATIÈRES

1 Préface 6

2 Introduction 7

3 Méthodologie d'enquête 8

4 Pondération longitudinale 14

4.1 Détermination des poids initiaux 14
4.2 Classification des individus longitudinaux 15
4.3 Modélisation et ajustement pour la non-réponse 16
4.4 Ajustement pour les valeurs influentes 21
4.5 Calage sur marges longitudinal 21
4.5.1 Théorie du calage 22
4.5.2 Application du calage sur marges pour les poids longitudinaux 23

5 Pondération transversale 26

5.1 Détermination des individus éligibles à la pondération transversale 26
5.2 Ajustement pour la non-réponse 27
5.3 Ajustement pour la migration interprovinciale 27
5.4 Application des facteurs d'allocation des panels 29
5.5 Partage des poids 33
5.5.1 Partage des poids pour le poids intégré 34
5.5.2 Partage des poids pour le poids travail 35
5.6 Ajustement pour les valeurs influentes 36
5.7 Calage sur marges 38

6 Changements à venir et développements futurs 42

7 Conclusion 43

Remerciements 44

Annexe A : Liste des variables utilisées dans le modèle d'ajustement pour la non-réponse 45

Bibliographie 46

1 Préface

Le premier document décrivant de façon exhaustive les pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu (EDTR) a été écrit par Isabelle Lévesque et Sarah Franklin et a paru en 2000. Depuis, plusieurs changements et nouveautés ont été introduits dans la méthodologie de pondération tant pour l'aspect longitudinal que transversal. Il s'est donc avéré nécessaire de faire une mise à jour de ce document. Voici un résumé des principaux changements apportés à la méthodologie de pondération et décrits dans ce nouveau document :

- En ce qui a trait à l'ajustement de non-réponse, un facteur d'ajustement est maintenant attribué aux poids des enfants et des personnes non-éligibles (voir sections 4.2 et 4.3)

- Dans la pondération transversale, l'ajustement pour la migration interprovinciale est maintenant effectué à la suite de l'ajustement de non-réponse. Auparavant, cette étape était effectuée à la suite du partage des poids (voir section 5.3)

- Des modifications ont été apportées à la méthodologie en ce qui a trait au partage des poids pour le poids travail. Cette étape est maintenant effectuée de façon à mieux préserver l'ajustement de non-réponse et donc le rapport entre les individus longitudinaux d'un ménage (voir section 5.5.2)

- Lors de la combinaison des panels dans la pondération transversale, il est nécessaire d'identifier les immigrants et de connaître leur date d'immigration afin de déterminer s'ils avaient une chance ou non d'être sélectionnés dans l'un ou l'autre des panels. Or, les questions sur l'immigration ne sont posées qu'aux personnes âgées de 16 ans et plus. Une stratégie d'imputation du statut d'immigration et de la date d'immigration pour les enfants de moins de 16 ans a été mise en place afin de permettre de leur assigner un meilleur facteur d'allocation des panels (voir section 5.4)

- Des modifications ont aussi été apportées en ce qui a trait à l'application du facteur d'allocation des panels. Le facteur p est maintenant appliqué seulement aux individus longitudinaux du panel le plus ancien qui étaient éligibles lors de la sélection du plus récent panel (voir section 5.4)

- L'étape de post-stratification a été remplacée par un calage sur marges suite à l'ajout de nouveaux totaux de contrôle. Pour la pondération longitudinale, un calage en deux étapes est maintenant effectué pour permettre l'utilisation de totaux de contrôle similaires à ceux utilisés pour la pondération transversale (voir les sections 4.5 et 5.7 pour le calage des pondérations longitudinale et transversale respectivement)

2 Introduction

L'Enquête sur la Dynamique du travail et du revenu (EDTR), introduite à l'année de référence 1993, est une enquête longitudinale par panel menée auprès des individus. Elle vise à mesurer les changements au niveau du bien-être économique des individus et les facteurs qui peuvent influencer ces changements, plus particulièrement les facteurs déterminants au niveau des caractéristiques démographiques, familiales et au niveau de l'activité. Le produit final de l'enquête est un ensemble de fichiers de microdonnées. Pour un aperçu général de l'enquête, se référer à Lavigne et Michaud (1998).

Originalement, l'objectif principal de l'EDTR visait à fournir des données longitudinales afin de produire des estimations et des analyses du point de vue longitudinal. L'aspect transversal de l'enquête est maintenant devenu tout aussi important que l'aspect longitudinal, surtout en raison de l'intégration de l'Enquête sur les finances des consommateurs (EFC), une enquête transversale, à l'EDTR (Cotton et coll., 1999). La combinaison de ces deux enquêtes s'est faite à partir de l'année de référence 1998. Pour plus de renseignements sur l'EFC, se référer à Statistique Canada (1997).

La pondération d'une enquête longitudinale peut parfois s'avérer un grand défi méthodologique, tant par ses aspects longitudinal et transversal, que par la nature dynamique de la composition du panel (engendrée directement par la dynamique des familles et des ménages au fil des ans). De nombreux défis ont été relevés au cours des années du point de vue de la pondération: la combinaison longitudinale de panels qui se chevauchent, l'ajout de nouveaux groupes de calage, la mise à jour des programmes de pondération afin de les rendre plus efficaces et versatiles, la recherche continue pour améliorer la méthodologie employée, etc. Tous ces éléments ajoutent à la complexité des pondérations longitudinale et transversale de l'EDTR.

Le but premier du présent document est de présenter les différentes étapes de la pondération, tant longitudinale que transversale, telles qu'utilisées lors de la production des poids pour l'année de référence 2003. L'autre objectif consiste à informer le lecteur des développements futurs et des changements importants à venir en ce qui concerne la pondération. Le document est divisé en cinq sections. Tout d'abord, la méthodologie de l'enquête est donnée à la section 3. Ensuite, les étapes de la pondération longitudinale, de la détermination des poids initiaux à l'obtention des poids longitudinaux finaux, sont décrites à la section 4. Puis, toutes les étapes de la pondération transversale sont discutées à la section 5. Il est à noter que pour la description de certaines étapes de la pondération, les aspects longitudinal et transversal sont intégrés. De plus, les considérations futures sont présentées à la section 6. Et finalement, la conclusion est donnée à la section 7.

Ce document est une mise à jour du document sur la pondération de l'EDTR, rédigé par Lévesque et Franklin (2000). Comme plusieurs changements et nouveautés ont été apportés à la méthodologie des pondérations longitudinale et transversale depuis, la rédaction d'un nouveau document s'est avérée nécessaire.

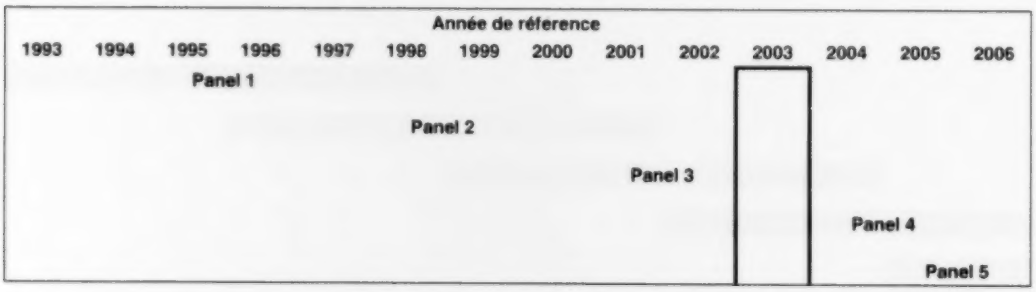
3 Méthodologie d'enquête

L'échantillon de l'EDTR est constitué de deux panels d'une durée de six ans chacun (ou six vagues). Chaque panel couvre la population des personnes habitant dans les dix provinces canadiennes à une date de référence précise. La population couverte par chacun des panels exclut les personnes résidant dans une réserve indienne, les membres à temps plein des Forces armées canadiennes habitant dans les casernements militaires ainsi que les pensionnaires dans les établissements institutionnels pour plus de six mois. Dans ce document, tout ce qui fera référence à la population des dix provinces canadiennes implique toutes ces exclusions.

Le premier panel a été sélectionné en janvier 1993 et couvre la population des dix provinces canadiennes en date du 31 décembre 1992. Le deuxième panel a été introduit en 1996 et depuis, un nouveau panel est sélectionné à tous les trois ans pour remplacer le panel le plus âgé. Ceci est fait dans le but d'améliorer la représentativité de l'échantillon transversal (pour tenir compte des nouveaux ménages qui se sont ajoutés à la population depuis trois ans), de diminuer l'effet de l'érosion et de réduire le fardeau de réponse. Par conséquent, il y a toujours un chevauchement de trois ans entre deux panels consécutifs.

La période utilisée dans ce document pour illustrer la stratégie de pondération, c'est-à-dire l'année de référence 2003, est couverte par le panel 3 et le panel 4 qui représentent la population des dix provinces canadiennes en date du 31 décembre des années 1998 et 2001 respectivement. La figure 3.1 donne un aperçu de la sélection des panels et du chevauchement entre ceux-ci.

Figure 3.1 Chevauchement des panels de l'EDTR

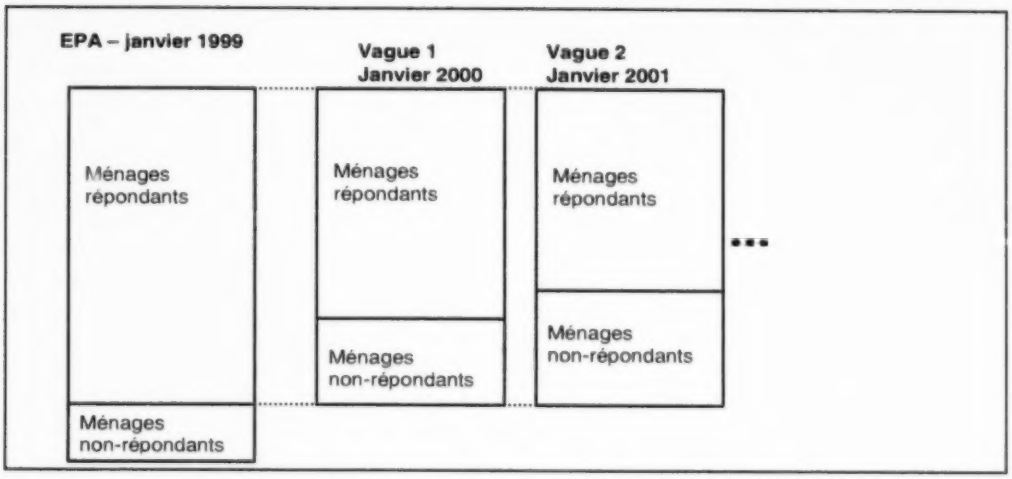


Chaque panel de l'EDTR consiste en un sous-échantillon d'environ 17 000 ménages (approximativement 40 000 personnes) provenant de l'Enquête sur la population active (EPA). L'échantillon de l'EPA est prélevé d'une base aréolaire selon un plan de sondage probabiliste à plusieurs degrés. L'EPA fonctionne sur une base de six panels (groupes de renouvellement) d'une durée de six mois chacun, un panel étant renouvelé à tous les mois. L'unité d'échantillonnage au dernier degré est le logement. Ainsi, tous les individus appartenant aux ménages qui occupent les logements choisis font partie de l'échantillon de l'EPA. Pour plus de renseignements sur l'EPA, se référer aux documents décrivant la méthodologie de cette enquête (Singh et coll., 1990; et Gambino et coll., 1998).

Les ménages sélectionnés pour l'EDTR sont ceux faisant partie des groupes de renouvellement sortant des panels de l'EPA aux deux premiers mois de la première période de référence de l'EDTR (soit en janvier et février 1999 pour le panel 3, et en janvier et février 2002 pour le panel 4). Parmi tous les ménages faisant partie des groupes de renouvellement sortant de l'EPA en janvier ou février, l'EDTR sélectionne seulement les ménages qui sont répondants au dernier mois de leur participation à l'EPA. La dernière entrevue de l'EPA sert alors de premier contact pour introduire l'EDTR et obtenir les coordonnées de personnes-contacts dans le but de faciliter le dépistage lors de la collecte de la première vague.

Lors de la première vague, une entrevue préliminaire est d'abord administrée à chaque personne sélectionnée par l'EDTR. Cette entrevue recueille des renseignements de base sur l'expérience de travail, les antécédents familiaux et personnels et le niveau d'instruction. Il est à noter que pour le premier panel, l'entrevue préliminaire a été menée en même temps que la dernière entrevue de l'EPA. Pour plus de détails sur les procédures suivies lors de l'introduction du premier panel, se référer à Lévesque et Franklin (2000). La figure 3.2 donne un aperçu de la sélection des ménages pour le panel 3 pour l'EDTR. Plus de détails sur la sélection des panels de l'EDTR sont donnés à la section 4.1.

Figure 3.2 Sélection des ménages pour l'échantillon longitudinal de l'EDTR du panel 3



Chaque personne sélectionnée pour l'EDTR est interviewée deux fois par année pendant six ans. À chaque année, des informations sont recueillies concernant la composition de la famille et du ménage, l'activité, l'éducation, le logement ainsi que sur les revenus de l'année précédente. En janvier, l'information sur l'activité, l'éducation, le logement est recueillie alors qu'en mai, on obtient l'information sur le revenu. La période de mai a été choisie pour faire la collecte de l'information sur le revenu parce que la plupart des individus ont alors rempli leur rapport d'impôt. Ceci permet de croire que l'information sur leurs revenus sera plus juste et complète à ce moment-là. Par ailleurs, afin de réduire le fardeau de réponse, les répondants peuvent éviter la collecte de mai en préautorisant Statistique Canada à utiliser leur dossier fiscal de l'Agence du revenu du Canada. Il est à noter que lors de la collecte de 2005 (pour l'année de référence 2004), l'entrevue de mai a été intégrée à celle de janvier de façon à réduire les coûts de collecte. Les impacts de ce changement sur la qualité des données sur le revenu ainsi que sur les taux de permission pour l'utilisation des dossiers fiscaux devront être évalués.

À la sélection d'un panel, toutes les personnes faisant partie des ménages choisis pour l'EDTR, indépendamment de l'âge de ces personnes, deviennent des membres de l'échantillon longitudinal du panel. Ces personnes sont considérées comme des membres de l'échantillon longitudinal pendant les six ans de participation au panel, même si elles déménagent, meurent, entrent en institution ou deviennent membres des forces armées à temps plein à l'intérieur des casernements militaires. Aucune autre personne ne devient membre de l'échantillon longitudinal pour ce panel. Ainsi, pour chaque panel, l'échantillon longitudinal est constitué au moment de l'introduction du panel et reste tel quel pendant toute la durée du panel. La population ciblée par l'échantillon longitudinal demeurera la même tout au long des six années de présence du panel à l'enquête.

Du point de vue longitudinal, l'unité d'intérêt est l'individu; le ménage pouvant difficilement servir d'outil d'analyse longitudinale à cause de sa nature dynamique dans le temps. Du point de vue transversal, on s'intéresse aussi bien à l'individu qu'au ménage. Comme l'EDTR s'intéresse non seulement aux

caractéristiques des personnes longitudinales mais aussi aux caractéristiques du ménage, on interviewe toutes les personnes habitant avec au moins un individu longitudinal. L'échantillon transversal pour une année donnée est alors constitué de toutes les personnes longitudinales dans le champ de l'enquête au 31 décembre de l'année de référence correspondante, et de toutes les personnes qui habitent avec elles à ce moment précis de l'année. Une personne est considérée dans le champ de l'enquête au 31 décembre d'une année de référence donnée (dans le champ de l'enquête transversalement) si elle habite dans une des dix provinces canadiennes, ne réside pas dans une réserve indienne ou un établissement institutionnel depuis plus de six mois, et n'est pas un membre à temps plein des Forces armées canadiennes habitant dans un casernement militaire, à ce moment de l'année. Toute personne longitudinale qui est dans le champ de l'enquête transversalement fera aussi partie de l'échantillon transversal. Toutes les personnes interviewées qui ne font pas partie de l'échantillon longitudinal sont appelées cohabitants. Il est à noter que l'ajout des cohabitants à l'échantillon transversal aide à améliorer la représentativité de l'échantillon transversal (Latouche et Michaud, 1997).

À chaque année, des poids longitudinaux et transversaux sont produits pour satisfaire ces besoins. La pondération longitudinale vise à produire des estimations qui sont représentatives de la population des dix provinces canadiennes au moment de la sélection de l'échantillon longitudinal alors que la pondération transversale permet de produire des estimations représentatives de la population des dix provinces canadiennes en date du 31 décembre d'une année de référence donnée. Par exemple, les estimations produites à partir de l'échantillon longitudinal du troisième panel sont représentatives de la population des dix provinces canadiennes au 31 décembre 1998 et celles produites avec l'échantillon longitudinal du quatrième panel font référence à la population des dix provinces canadiennes au 31 décembre 2001. Les estimations produites à partir de l'échantillon transversal pour l'année de référence 2003 représentent la population des dix provinces canadiennes au 31 décembre 2003.

En janvier 1999, 43 683 personnes ont été sélectionnées pour faire partie de l'échantillon longitudinal du troisième panel et en janvier 2002, 42 232 personnes ont été choisies pour faire partie de l'échantillon longitudinal du quatrième panel. En ce qui a trait à l'échantillon transversal pour l'année de référence 2003, celui-ci comprend 75 206 individus répondants, dont 65 329 personnes longitudinales.

Le tableau 3.1 donne un aperçu des changements dans la constitution des échantillons longitudinaux au cours du temps pour les panels 3 et 4 alors que le tableau 3.2 présente la constitution des échantillons transversaux pour les années de référence 1999 à 2003 inclusivement.

Tableau 3.1 Constitution des échantillons longitudinaux des panels 3 et 4 pour les années de référence 1999 à 2003

	Année de référence				
	1999	2000	2001	2002	2003
	Panel 3 - Personnes				
Répondants[1]	nombre				
Dans le champ de l'enquête au 31 décembre d'une année de référence donnée[2]	36 158	35 340	34 892	32 922	31 214
Habitant en dehors des dix provinces canadiennes	130	234	336	400	377
En institution	97	169	255	437	514
Décédée	274	496	761	1 010	1 283
Duplicata, erreur	0	0	0	3	9
Non-répondants[3]					
Refus ferme, personnes dans les ménages non dépistés	7 024	7 444	7 439	8 911	10 286
Total	43 683	43 683	43 683	43 683	43 683
	Panel 4 - Personnes				
Répondants[1]	nombre				
Dans le champ de l'enquête au 31 décembre d'une année de référence donnée[2]	...	...	...	33 792	34 115
Habitant en dehors des dix provinces canadiennes	...	...	...	108	241
En institution	...	...	...	141	265
Décédée	...	...	...	255	515
Duplicata, erreur	...	...	...	0	14
Non-répondants[3]					
Refus ferme, personnes dans les ménages non dépistés	...	...	...	7 936	7 082
Total	...	...	...	42 232	42 232

0 zéro absolue ou valeur arrondie à zéro
... n'ayant pas lieu de figurer
1. Les individus longitudinaux répondants auront un poids longitudinal non nul.
2. Habitant toujours dans les dix provinces canadiennes, n'est pas décédé, ne réside pas dans un établissement institutionnel depuis plus de six mois.
3. Les individus longitudinaux non-répondants auront un poids longitudinal égal à 0. Cette catégorie comprend les cas de refus ferme ou non contactés durant la collecte ainsi que les cas qui ne sont pas renvoyés à la collecte car ils ont été non-répondants pour deux années consécutives.

Tableau 3.2 Constitution des échantillons transversaux, par panel pour les années de référence 1999 à 2003

	Année de référence				
	1999	2000	2001	2002	2003
	Panel 2 – Personnes				
Répondants	nombre				
Personnes longitudinales	34 426	32 224	31 451	...	...
Cohabitants	5 673	6 210	7 306	...	...
Sous-total	40 099	38 434	38 757	...	...
	Panel 3 – Personnes				
Répondants	nombre				
Personnes longitudinales	36 158	35 340	34 892	32 922	31 214
Cohabitants	1 869	3 072	4 830	5 753	6 406
Sous-total	38 027	38 412	39 722	38 675	37 620
	Panel 4 – Personnes				
Répondants	nombre				
Personnes longitudinales	...	...	...	33 792	34 115
Cohabitants	...	...	...	2 012	3471
Sous-total	...	...	...	35 804	37 586
Total - Répondants transversaux[1]	78 126	76 846	78 479	74 479	75 206

... n'ayant pas lieu de figurer
1. Ces individus auront un poids transversal non-nul pour l'année de référence correspondante.

Il est à noter que lorsqu'il s'agit de la constitution des échantillons transversaux (tableau 3.2), le groupe des non-répondants a peu d'importance puisque l'ajustement de non-réponse n'est fait qu'à partir de l'échantillon des longitudinaux (pour plus de détails à propos de l'ajustement de non-réponse pour les pondérations longitudinale et transversale, voir les sections 4.3 et 5.2). Par ailleurs, il est difficile d'avoir une juste représentation des cohabitants non-répondants puisqu'il est souvent impossible de savoir la composition actuelle d'un ménage non-répondant. Il est possible qu'un cohabitant ne fasse plus partie du ménage ou que d'autres cohabitants s'y soient ajoutés sans qu'on puisse le savoir.

L'EDTR a pour mandat de fournir des données longitudinales et transversales sur les caractéristiques reliées à l'activité et au revenu, et de rendre ces données disponibles aux utilisateurs aussi bien à l'intérieur qu'à l'extérieur de Statistique Canada. Pour ce faire, un ensemble de fichiers de microdonnées non traitées pour la confidentialité (fichiers internes) et un ensemble de fichiers de microdonnées traitées pour la confidentialité (fichiers externes) sont produits à partir de l'EDTR.

Après le traitement des données d'une année de référence donnée, les utilisateurs de Statistique Canada peuvent avoir accès à la base de données relationnelle pour faire leurs analyses transversales et longitudinales. Toute information sur cette base de données est confidentielle étant donné qu'elle n'a pas reçu de traitement pour la confidentialité. Un système d'extraction de données a aussi été développé afin de

faciliter l'accès aux données de l'enquête pour les utilisateurs de Statistique Canada (\\lhs2\clientserv\Income\SLID\SLIDCopy\SLIDRet_Installation).

Les utilisateurs externes peuvent avoir accès aux fichiers de microdonnées externes qui sont généralement produits pour chaque année de référence. Plus de travail est nécessaire pour diffuser de tels fichiers puisqu'ils doivent être traités pour la confidentialité. Entre autres, l'ensemble des mesures de contrôle de la divulgation est évalué et amélioré lors de chaque diffusion pour assurer la confidentialité des données fournies. Pour les périodes de référence 1993 et 1994, des fichiers de microdonnées longitudinaux pour les besoins externes ont été produits (voir Lévesque et Franklin, 2000). Cependant, en raison des difficultés à préserver la confidentialité des données pour de tels fichiers, seuls les fichiers transversaux de microdonnées ont été diffusés par la suite. La stratégie de diffusion consiste maintenant à produire un fichier transversal de microdonnées à grande diffusion pour chaque année de référence. Afin de réduire les risques de divulgation associés à la diffusion de tels fichiers, les fichiers de microdonnées sont maintenant constitués d'un sous-échantillon de ménages tiré à partir de l'échantillon transversal original. Une repondération est effectuée de façon à ce que le sous-échantillon soit représentatif de la population cible pour une période de référence donnée.

Les fichiers transversaux de microdonnées (faisant référence à l'année de référence traitée) comprennent donc un sous-échantillon de toutes les personnes (individus longitudinaux et cohabitants) dans les ménages répondants et qui sont dans le champ de l'enquête au 31 décembre de l'année de référence ciblée.

4 Pondération longitudinale

Cette section a pour but de décrire la méthodologie utilisée pour pondérer à chaque vague les individus longitudinaux de l'EDTR. Afin de mieux illustrer la méthodologie utilisée, des exemples tirés de la cinquième vague du troisième panel et de la deuxième vague du quatrième panel sont présentés. Il est important de rappeler que les poids longitudinaux visent à représenter la population des dix provinces canadiennes au 31 décembre 1998 pour le troisième panel et au 31 décembre 2001 pour le quatrième panel. Ainsi, les ensembles de poids longitudinaux discutés dans le présent document sont produits séparément pour les deux panels[1]. Il est à noter que depuis 2002, un nouvel ensemble de poids longitudinal combinant deux panels se chevauchant est aussi produit. Pour plus de détails sur la méthodologie employée pour la création de cet ensemble de poids, prière de se référer à Naud (2004).

La pondération longitudinale s'effectue selon les cinq étapes présentées aux sections suivantes : la détermination des poids initiaux (section 4.1), la classification des individus longitudinaux (section 4.2), la modélisation et l'ajustement pour la non-réponse (section 4.3), l'ajustement pour les valeurs influentes (section 4.4) et enfin le calage pour ajuster certaines estimations de l'enquête à des totaux connus de la population (section 4.5).

4.1 Détermination des poids initiaux

Pour le troisième échantillon de l'EDTR (panel 3), tous les ménages qui ont répondu à l'EPA en janvier et février 1999 pour les groupes de rotation sortant de l'EPA en janvier et février 1999 respectivement, ont été sélectionnés (17 085 ménages pour un total de 43 683 personnes). Les entrevues préliminaires du troisième panel ont été menées en même temps que les entrevues de la première vague (c'est-à-dire en janvier 2000), comme ce fût le cas à partir de l'introduction du deuxième panel[2].

La sélection de l'échantillon du quatrième panel de l'EDTR s'est effectuée de la même manière que celle du panel 3. Ainsi tous les ménages qui ont répondu à l'EPA en janvier et février 2002 pour les groupes de rotation sortant de l'EPA en janvier et février 2002 respectivement, ont été sélectionnés (16 942 ménages pour un total de 42 232 personnes).

Les poids initiaux de la pondération longitudinale du troisième et du quatrième panel correspondent à l'inverse de la probabilité de sélection du ménage. Ainsi tous les membres longitudinaux d'un même ménage (celui auquel ils appartenaient lors de la sélection de l'échantillon) ont le même poids initial.

L'échantillon de l'EDTR étant tiré de l'échantillon de l'EPA, le poids initial sera donc :

$$W_{initial\,.\,p_i} = W_{EPA} * 3$$

où $W_{initial\,.\,p_i}$ = Poids longitudinal initial pour le panel *i, i=3,4*

W_{EPA} = Poids de l'EPA ajusté pour la non-réponse (sous-poids de l'EPA)

3 = Inverse de la probabilité de sélection des groupes de rotation de l'EPA (2 groupes sur 6 ont été sélectionnés)

---

1. Auparavant, l'EDTR produisait deux ensembles de poids longitudinaux : un ensemble de poids internes et un ensemble de poids externes. Les poids destinés à l'externe n'étant plus utilisés, seul un ensemble de poids internes est maintenant produit. Pour plus de détails sur la création des anciens poids externes, se référer à Lévesque et Franklin (2000).
2. Les procédures de sélection du panel 1 ayant été différentes, se référer à Lévesque et Franklin (2000) pour plus de détails sur la détermination des poids initiaux propres à ce panel.

Le tableau 4.1 donne les médianes des poids longitudinaux initiaux par province au moment de la sélection des panels, c'est-à-dire la médiane des $w_{initial,p_i}$, pour i=3, 4.

Tableau 4.1 Médianes des poids longitudinaux initiaux, pour le panel 3 et le panel 4

Province	Longitudinaux initiaux	
	Panel 3	Panel 4
	poids	
Terre Neuve et Labrador	237	238
Île du Prince Édouard	101	103
Nouvelle-Écosse	288	279
Nouveau-Brunswick	279	265
Québec	567	561
Ontario	560	566
Manitoba	336	329
Saskatchewan	280	282
Alberta	692	796[1]
Colombie-Britannique	843	834

1. La différence plus importante pour l'Alberta pourrait être causée par un certain nombre de grappes à forte croissance qui seraient entrées ou sorties de l'échantillon de l'EPA durant cette période.

4.2 Classification des individus longitudinaux

À chaque vague, la première étape de la pondération longitudinale est de classer les individus longitudinaux parmi les catégories suivantes : les non-répondants (dans les ménages répondants ou non-répondants), les répondants qui sont dans le champ de l'enquête transversalement et les personnes qui sont hors du champ de l'enquête transversalement. Un ménage répondant est un ménage dans lequel il y a au moins une personne qui a répondu à au moins une des deux entrevues (l'entrevue sur le travail ou l'entrevue sur le revenu). Un ménage non-répondant est un ménage dans lequel tous les membres du ménage sont non-répondants aux deux entrevues. Il a été déterminé en outre que les renseignements sur la composition du ménage ne peuvent être manquants pour plus d'un an. Si c'est le cas, alors le ménage sera considéré comme non-répondant. Une personne est considérée hors du champ de l'enquête transversalement (ou non-éligible transversalement) si, au 31 décembre de l'année de référence, elle était décédée, en institution pour plus de six mois ou ne résidait pas dans une des dix provinces canadiennes.

La classification des individus longitudinaux permet de déterminer quelles personnes doivent recevoir un poids longitudinal non nul. Un poids longitudinal nul n'est attribué qu'aux individus dans les ménages non-répondants. Tous les autres individus qui étaient sélectionnés initialement pour faire partie de l'échantillon longitudinal ont un poids longitudinal non nul, même s'ils sont hors du champ de l'enquête transversalement (puisqu'ils sont considérés comme étant répondants). Il est à noter que du point de vue de la pondération, l'EDTR définit les personnes non-répondantes dans les ménages répondants comme répondantes. Cela est fait pour assurer la cohérence intra-ménage en ce qui a trait aux poids (les données des personnes non-répondantes dans les ménages répondants sont alors imputées).

La classification des individus longitudinaux permet aussi de déterminer si le poids d'un individu devrait être ajusté pour la non-réponse, en plus d'identifier les individus qui seront utilisés dans le modèle d'ajustement pour la non-réponse. Ce modèle sert à trouver un ajustement adéquat pour gonfler les poids des répondants

afin que ces derniers soient représentatifs de l'ensemble des répondants et des non-répondants. Des changements ont été apportés à la méthodologie de non-réponse lors de la révision historique de 2005. Ceux-ci ont été appliqués rétroactivement pour l'ensemble des années de l'EDTR et touchent plus particulièrement les enfants dans les ménages répondants et ainsi que les personnes étant hors du champ de l'enquête. Un bref aperçu des changements apportés est présenté dans les paragraphes suivants. Cependant, pour de plus amples renseignements sur l'ancienne méthodologie d'ajustement de la non-réponse, on peut se référer à Lévesque et Franklin (2000).

Dans l'ancienne méthodologie, le poids des enfants et des personnes hors cible n'était pas ajusté pour compenser pour la non-réponse. En ce qui concerne le poids des enfants, cette décision était justifiée par le fait qu'il y a trop peu d'information sur ces derniers pour permettre de modéliser adéquatement la non-réponse et aussi que l'ajustement fait lors de la stratification *a posteriori* ajustait pour la non-réponse de façon implicite. Or, lors d'études sur le glissement de l'EDTR, les résultats ont démontré que le fait de ne pas ajuster le poids des enfants pour la non-réponse nuisait à la représentation des ménages avant l'étape de calage. En effet, les ménages ayant un nombre plus élevé d'enfants étaient beaucoup moins bien représentés, ce qui engendrait par la suite une plus grande distorsion des poids à l'étape du calage. Étant donné le peu d'information disponible sur les enfants et que la décision de répondre ou non à l'enquête ne leur revient généralement pas, ces derniers sont exclus de la modélisation de la non-réponse. Cependant, ils reçoivent quand même un ajustement de non-réponse équivalent à l'ajustement moyen des adultes longitudinaux vivant dans le ménage.

En ce qui a trait aux personnes hors du champ de l'enquête, il s'est avéré que ceux-ci ne sont pas surreprésentés contrairement à ce qui avait été présumé. En effet, de nouvelles analyses ont démontré que cette population était plutôt sous-représentée (Godbout, 2004). Ainsi, ces résultats et le fait que certains non-répondants sont probablement hors du champ de l'enquête, ont justifié la ré-inclusion de ces derniers dans l'ajustement de non-réponse. Cependant, ces derniers sont exclus de l'étape de la modélisation de la non-réponse.

Un ajustement pour la non-réponse est habituellement appliqué aux poids de toutes les unités répondantes de l'enquête. Dans le cas de l'EDTR, un facteur d'ajustement pour la non-réponse est appliqué aux poids de toutes les personnes considérées comme répondantes.

Le tableau 4.2 présente les différentes catégories de réponse des individus longitudinaux pour la cinquième vague du troisième panel et la deuxième vague du quatrième panel. La colonne *Modélisation de la non-réponse* indique quelles catégories d'individus sont utilisées dans le modèle d'ajustement pour la non-réponse. Les deux dernières colonnes indiquent quel traitement sera appliqué aux poids des individus faisant partie des différentes catégories.

4.3 Modélisation et ajustement pour la non-réponse

Une fois les poids initiaux calculés, il faut maintenant calculer un facteur d'ajustement pour la non-réponse pour chaque individu considéré comme répondant. Rappelons ici que les deux panels sont traités séparément. Le facteur d'ajustement pour la non-réponse appliqué à un individu adulte est défini comme étant l'inverse du taux de réponse du groupe homogène de réponse (GHR) dans lequel celui-ci se situe ($1/R_{GHR}$). Les GHR sont formés en regroupant ensemble des individus qui semblent être affectés par la non-réponse de la même manière. Si les GHR sont définis de telle sorte que la non-réponse soit complètement aléatoire à l'intérieur d'un groupe, alors le biais dû à la non-réponse est négligeable (Tambay et coll., 1998). Pour trouver un ajustement adéquat pour gonfler les poids des répondants afin que ces derniers soient représentatifs de l'ensemble des répondants et des non-répondants, on procède à la modélisation de la non-réponse.

La première étape de la modélisation de la non-réponse consiste à déterminer quelles sont les variables reliées à la non-réponse qui doivent être considérées dans le modèle afin de bien prédire la variable dichotomique *réponse* (dont les modalités sont simplement *répondant* ou *non-répondant*). Il y a deux façons possibles de procéder en ce qui a trait à l'ajustement de non-réponse. L'ajustement peut être fait en fonction des caractéristiques initiales des individus faisant partie de l'échantillon ou en fonction des caractéristiques disponibles les plus récentes. Il existe des avantages et désavantages aux deux méthodes.

Tableau 4.2 Classification des individus longitudinaux pour la cinquième vague du troisième panel et la deuxième vague du quatrième panel

Classification des individus	Panel 3 vague 5	Panel 4 vague 2	Inclus dans la modélisation de la non-réponse	W_{ajust}
Enfants âgés de 0 à 15 ans (au 31 déc. 2003)				
Dans un ménage non-répondant	**1 538**	**1 315**	Non	**= 0**
Dans un ménage répondant	**4 931**	**6 420**	Non	$W_{ajust} = \frac{W_{initial}}{\overline{R}_{GHR,\ adultes}}$
Non-éligible transversalement (en institution, décédé ou en dehors des 10 provinces)	**58**	**28**	Non	$W_{ajust} = \frac{W_{initial}}{R_{GHR}}$
Adultes âgés de 16 ans et plus (au 31 déc. 2003)				
Non-éligibles transversalement (en institution, décédé ou en dehors des 10 provinces)	**2 116**	**993**	Non	$W_{ajust} = \frac{W_{initial}}{R_{GHR}}$
Répondants	**25 884**	**26 957**	Oui	$W_{ajust} = \frac{W_{initial}}{R_{GHR}}$
Non-répondants dans un ménage non-répondant	**8 757**	**5 781**	Oui	**= 0**
Non-répondants dans un ménage répondant (seront imputés)	**399**	**738**	Non	$W_{ajust} = \frac{W_{initial}}{R_{GHR}}$
Total	**43 683**	**42 232**	...	...

... n'ayant pas lieu de figurer

R_{GHR} : Taux de réponse dans le groupe homogène de réponse (GHR)

$\overline{R}_{GHR,\ adultes}$: Taux de réponse moyen des adultes de 16 ans et plus du ménage dans lequel l'enfant habite.

La première méthode utilise l'information recueillie par l'EPA, c'est-à-dire celle obtenue avant les premières entrevues de l'EDTR. Le principal avantage de cette méthode est que l'information utilisée est disponible pour tous les individus longitudinaux, qu'ils soient répondants ou non à la vague en traitement. Par contre, le principal inconvénient est que certaines des variables recueillies sont de nature dynamique (comme le revenu, le niveau d'éducation, l'état matrimonial, l'emploi, la présence d'enfants, etc.) et donc l'information risque d'être désuète à la date de référence de la vague traitée. Ainsi, faire un ajustement en utilisant l'information provenant de la vague initiale consiste davantage à réajuster l'échantillon de répondants de façon à ce qu'il représente bien l'échantillon initial plutôt qu'à un ajustement de non-réponse typique puisque les variables utilisées ne sont peut-être pas celles qui expliquent le mieux pourquoi un ménage ou un individu a répondu ou non à l'enquête. Cette méthode est celle qui a été privilégiée et utilisée jusqu'à présent par l'EDTR. En effet, l'EDTR ne considère que l'information connue obtenue avant le début de la collecte des données de l'enquête, c'est-à-dire celle qui a été recueillie lors de la dernière entrevue de l'EPA[3].

La deuxième méthode consiste à utiliser les renseignements connus les plus à jour pour les individus. Cette façon de procéder comporte aussi des risques. D'abord, les informations utilisées peuvent ne pas être issues de la même période. En effet, les renseignements des ménages longitudinaux répondants peuvent provenir

3. Il est à noter que pour le panel 1, l'information utilisée était celle recueillie lors de l'interview préliminaire étant donné que celle-ci a eu lieu en même temps que la dernière entrevue de l'EPA (Lévesque et Franklin, 2000).

de la vague courante alors que ceux des ménages non-répondants peuvent provenir de la vague antérieure. De plus, en considérant l'information la plus à jour pour la modélisation de la non-réponse, il faudrait utiliser le poids associé à la vague précédant la vague en traitement comme poids de départ au lieu d'utiliser le poids initial de l'EDTR. Or les non-répondants convertis, c'est-à-dire des individus qui étaient non-répondants à la vague antérieure et qui deviennent répondants à la vague en traitement posent un certain problème. Comme ces individus convertis ont un poids longitudinal nul à la vague précédant celle en traitement, il faudrait trouver une façon de leur assigner un poids de départ.

Une étude a été menée dans le but d'évaluer la possibilité de modifier la méthodologie employée par l'EDTR pour l'ajustement de non-réponse (Naud, 2004). Celle-ci visait principalement à tester l'utilisation d'informations provenant des vagues les plus récentes dans la modélisation de la non-réponse. Deux nouvelles méthodes d'ajustement ont été développées et évaluées. La conclusion tirée de cette étude est que l'utilisation des données provenant de vagues plus récentes dans la modélisation de la non-réponse comporte certains avantages mais il n'a pas été possible de démontrer qu'un tel changement menait clairement à des estimations de meilleure qualité. Ainsi, étant donné les résultats mitigés obtenus et la complexité des nouvelles méthodes proposées, il a été décidé de ne pas changer la méthodologie actuelle et de poursuivre les études sur les différentes méthodes possibles pour améliorer la modélisation de la non-réponse.

Tel que mentionnée ci-haut, les informations provenant de l'EPA sont celles utilisées pour la modélisation de la non-réponse de l'EDTR. Comme il y a plusieurs variables disponibles provenant de l'EPA, la première étape est de déterminer quelles variables indépendantes seront utilisées pour la modélisation de la non-réponse. Cette première étape s'effectue seulement lors du traitement de la première vague d'un panel puisque les variables considérées dans le modèle de non-réponse sont fixées au début de chaque panel. Aux vagues subséquentes, ces mêmes variables choisies sont aussi utilisées. Pour déterminer les variables explicatives à considérer pour la modélisation de la non-réponse, les variables catégoriques dont le nombre de modalités (m) est supérieur à deux sont recodées en m variables dichotomiques. À partir de toutes ces variables, on établit subjectivement une liste des variables les plus susceptibles d'influencer la réponse.

Pour chacune des variables de cette liste, des régressions logistiques sont produites en prenant une variable à la fois. D'après les résultats de ces régressions logistiques univariées, les variables qui sont significatives sont retenues. En tout, 46 variables dichotomiques ont été conservées pour les troisième et quatrième panels. La liste complète de ces variables est fournie par panel à l'annexe A. Comme mentionné précédemment, les mêmes variables sont conservées d'une vague à l'autre pour un même panel; les seuls changements habituellement effectués sont des modifications aux modalités de certaines variables (comme les variables de groupe d'âge et de revenu).

La deuxième étape est de dériver les GHR en considérant les variables retenues à l'étape précédente. L'EDTR utilise la méthode de modélisation par segmentation pour déterminer les GHR[4]. Avant l'année de référence 2002, l'EDTR se servait du logiciel Knowledge Seeker pour dériver les GHR. Depuis, l'équipe de l'EDTR a développé un programme en langage SAS, qui permet de dériver les GHR de la même façon que ce que fait le logiciel Knowledge Seeker, mais d'une façon automatisée et plus conviviale puisqu'aucune intervention manuelle n'est nécessaire.

La modélisation par segmentation utilise un processus itératif pour partitionner le fichier de données et ainsi former l'arbre de décision. Le premier noeud de l'arbre de décision est tout simplement l'ensemble complet des données. La première itération définit le premier embranchement de l'arbre. Ce dernier est créé en déterminant la variable qui influence le plus significativement la variable réponse parmi les variables retenues pour la modélisation. Une mesure statistique reflétant l'importance de la variable pour discriminer la non-réponse est calculée pour chaque variable; d'après cette mesure, la variable qui est la plus significativement importante est choisie. Le premier embranchement est ainsi formé, créant un noeud à chaque extrémité. Pour chaque noeud formé, une nouvelle itération est effectuée en répétant le même mécanisme de façon

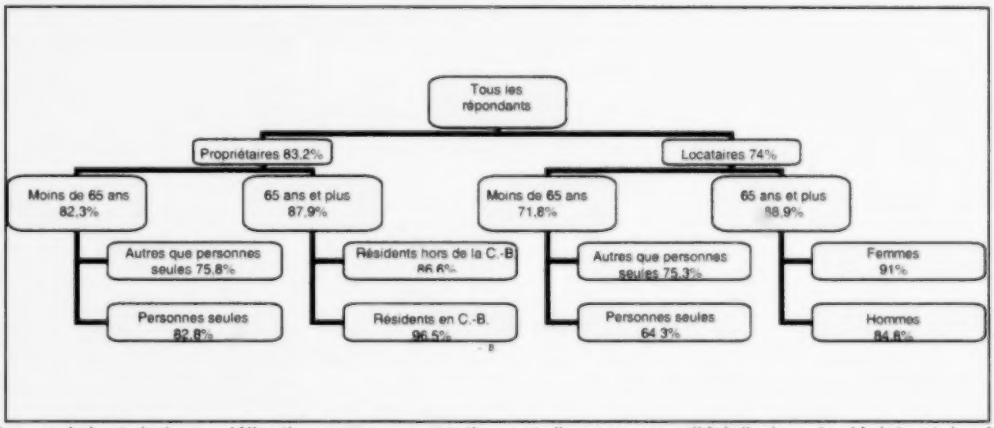
4. Cette méthode a été préférée à la régression logistique puisqu'elle semblait plus efficace pour réduire le biais dû à la non-réponse selon Dufour et coll., 1998. Cependant, certaines enquêtes semblent maintenant privilégier la régression logistique et de nouvelles études pourraient être entreprises pour évaluer si la méthode de modélisation par segmentation utilisée devrait être changée.

indépendante. De nouveaux embranchements et de nouveaux noeuds sont alors créés. Si, pour un noeud donné, aucune variable n'est significative pour expliquer la réponse, alors aucun autre embranchement n'est créé. La formation de cette branche de l'arbre est alors complétée. Le processus itératif se termine lorsque toutes les branches de l'arbre sont achevées.

Plusieurs méthodes peuvent être employées pour déterminer la variable qui influence le plus la propension à répondre. L'EDTR utilise l'algorithme CHAID ("Chi-Square Automatic Interaction Detection") qui sélectionne la variable avec la valeur de la statistique du khi-deux de Pearson la plus élevée.

Une des raisons pour laquelle la modélisation par segmentation est préférable à la régression logistique est qu'elle est plus flexible. La régression logistique force le modèle à être symétrique : une fois que les variables significatives sont déterminées, les GHR sont formés en croisant l'ensemble de ces variables. En forçant le modèle à être symétrique, de petites cellules ou des cellules qui ne sont pas significatives peuvent être formées, ce qui peut affecter la fiabilité des estimations. La modélisation par segmentation permet au modèle d'être asymétrique : à chaque embranchement de l'arbre de décision, on détermine quelle variable est la plus significative; ce qui garantie que chaque GHR formé est significatif et résulte habituellement en un modèle asymétrique. La figure 4.1 montre les variables les plus importantes pour discriminer la réponse et la non-réponse dans le cas du panel 4, pour l'année de référence 2003. Ainsi, la variable la plus importante dans ce cas est une variable indiquant si oui ou non une personne est propriétaire d'une maison.

Figure 4.1 Premières branches de l'arbre de décision pour le panel 4 (taux de réponse), année de référence 2003



Un inconvénient de la modélisation par segmentation est directement relié à l'arbre de décision lui-même. Lorsqu'un arbre de décision est formé, il peut facilement devenir assez lourd : dès qu'il comporte plusieurs embranchements, il devient difficile à interpréter. Les arbres de décision sont aussi assez instables; l'ajout, le retrait ou le changement d'une variable ou d'un individu peut avoir un effet important sur l'arbre lui-même. Il est alors très difficile de comparer plusieurs arbres de décision entre eux.

À chaque vague, l'EDTR crée un modèle d'ajustement pour la non-réponse séparément pour chacun des panels. Au départ, chaque modèle est défini au niveau national. Des embranchements au niveau provincial sont retenus dans l'arbre de décision si ces derniers sont significatifs. L'avantage de procéder à l'ajustement de non-réponse à un niveau national est que ceci permet d'utiliser un plus grand nombre de variables pour la création des GHR[5].

5. Des études ont démontré que cette approche engendre des différences entre les effectifs provinciaux avant et après l'ajustement de non-réponse. Plus particulièrement, les effectifs des petites provinces sont souvent augmentés alors que ceux des plus grosses provinces sont réduits. D'autres études sont prévues dans le but d'étudier plus à fond cette problématique et déterminer s'il n'aurait pas lieu de faire un ajustement de non-réponse au niveau provincial ou encore régional.

Pour tenir compte du plan d'échantillonnage, les poids initiaux normés sont utilisés pour calculer une statistique du khi-deux pondérée. Il est à noter que pour l'instant, l'EDTR ne fait pas de distinction entre les différents types de non-réponse (non-réponse causée par l'érosion, non-dépisté, refus, etc.)[6]. Il est également important de mentionner que certaines contraintes sont appliquées au modèle lors de la modélisation par segmentation. En effet, chaque GHR doit contenir au moins trente personnes (pondéré et non pondéré) et doit avoir un taux de réponse pondéré d'au moins 50%. Kalton et Kasprzyk (1986) suggèrent d'avoir de telles contraintes puisqu'elles permettent d'éviter une trop grande variabilité dans les poids et diminuent la perte de précision dans les estimations.

Une fois l'arbre de décision obtenu, les GHR sont formés en considérant chaque extrémité des branches de l'arbre. Les individus qui ont été utilisés pour modéliser la non-réponse (i.e. les adultes répondants ainsi que les adultes non-répondants dans les ménages non-répondants) sont alors classés dans les différents GHR. Il est à noter que les adultes non-répondants dans les ménages répondants ne sont pas utilisés lors de la modélisation de la non-réponse même s'ils sont considérés comme répondants puisque dans les faits ils n'ont pas répondu à l'enquête.

L'ajustement pour la non-réponse est fait à l'intérieur de chaque GHR : le poids des répondants adultes d'un GHR est divisé par le taux de réponse pondéré observé dans le GHR (ce qui revient à multiplier le poids des répondants par un facteur d'ajustement correspondant à l'inverse du taux de réponse pondéré observé). Il est à noter que les personnes non-éligibles (en institution, décédées ou en dehors des dix provinces) qui ont été exclues de la modélisation sont incluses dans le calcul des taux de réponse de chaque GHR. Pour la cinquième vague du troisième panel, 124 GHR ont été formés et avaient un taux moyen de réponse pondéré de 71.8 %; alors que 209 GHR ont été formés pour la deuxième vague du quatrième panel, avec un taux moyen de réponse pondéré de 80.8 %.

Tel que mentionné précédemment, les enfants ne sont pas inclus dans la modélisation de la non-réponse. Cependant, leurs poids sont quand même ajustés pour la non-réponse de façon à ne pas créer un déséquilibre entre la représentation des ménages avec et sans enfant. Ainsi, le poids des enfants dans les ménages répondants est divisé par la moyenne des taux de réponse pondérés dans les GHR des adultes habitant le même ménage. Le facteur d'ajustement pour la non-réponse des enfants est équivalent à l'inverse de la moyenne des ajustements pour les adultes habitant le même ménage.

Finalement, les poids des non-répondants sont mis à 0. Ainsi, à chaque vague, les poids longitudinaux ajustés pour la non-réponse sont obtenus de la façon suivante :

$$w_{ajust} = \begin{cases} 0, & \text{Pour les personnes dans les ménages non-répondants} \\ \dfrac{w_{initial}}{R_{GHR}}, & \text{Pour les personnes adultes dans les ménages répondants et les personnes hors champ} \\ \dfrac{w_{initial}}{\overline{R}_{GHR\ adultes}}, & \text{Pour les enfants dans les ménages répondants} \end{cases}$$

où w_{ajust} = poids longitudinal ajusté pour la non-réponse

$w_{initial}$ = poids longitudinal initial (= $w_{initial,p_3}$ pour le panel 3 ou $w_{initial,p_4}$ pour le panel 4)

R_{GHR} = taux de réponse pondéré observé dans le GHR auquel la personne adulte appartient

$\overline{R}_{GHR,adultes}$ = moyenne des taux de réponse pondérés observés dans les GHR des adultes du ménage dans lequel l'enfant habite

6. Des études futures sont prévues pour analyser la possibilité de modéliser la non-réponse en tenant compte des différents types de non-réponse.

4.4 Ajustement pour les valeurs influentes

Une fois que les poids ont été ajustés pour compenser pour la non-réponse, on procède à l'identification des observations influentes qui ont un impact important sur les estimations ponctuelles du revenu ainsi que sur leur estimation de variance. Une valeur est dite influente si elle contribue trop fortement à l'estimation transversale du total des revenus provinciaux. Un facteur d'ajustement entre 0 et 1 est calculé et multiplié aux poids des individus identifiés comme étant influents (ainsi que le poids des autres membres de leur ménage) afin de diminuer la contribution des poids de ces individus sur les estimations du revenu. Pour la cinquième vague du troisième panel, 10 individus ont été identifiés comme étant influents, et pour la deuxième vague du quatrième panel, 22 individus ont été identifiés comme étant influents.

L'identification des valeurs influentes est seulement effectuée par rapport aux estimations transversales. La même procédure de calcul d'ajustement pour les valeurs influentes pourrait s'appliquer également avec les poids longitudinaux. On obtiendrait ainsi des ajustements pour les valeurs influentes potentiellement différents pour les pondérations transversale et longitudinale. Cependant, des tests faits pour l'année de référence 1995 ont montré qu'à ce moment-là, il n'y aurait pas eu de différence quant aux individus identifiés comme influents, et très peu de différence dans les facteurs d'ajustement comme tels. Il a donc été décidé d'utiliser, pour la pondération longitudinale, les mêmes enregistrements identifiés comme étant influents du côté transversal. Ainsi, l'identification des valeurs influentes est la même pour les aspects longitudinal et transversal. Pour plus de détails sur l'identification des valeurs influentes et sur le calcul du facteur d'ajustement à apporter aux poids, se référer à la section 5.6.

En ce qui a trait au calcul même du facteur d'ajustement, certaines modifications ont été apportées pour l'aspect longitudinal. Le but de ce changement a été de s'assurer que le poids d'une observation jugée influente à une vague *t* soit aussi ajusté pour les vagues subséquentes. Ceci a été fait parce que le poids utilisé dans les analyses longitudinales est souvent celui de la dernière vague disponible et que la combinaison de ce poids avec des données de revenu des vagues précédentes peut avoir trop d'influence sur les estimations. Si l'ajustement est appliqué seulement lors du traitement de la vague pour laquelle l'observation influente a été observée, le poids d'une observation pourrait ne pas être ajusté pour une année spécifique mais tout de même engendrer une estimation influente lorsque celui-ci est combiné avec le revenu d'une année antérieure. Donc pour l'aspect longitudinal, l'ajustement appliqué pour une observation jugée influente correspond au minimum des facteurs d'ajustements calculés pour la vague courante et l'ensemble des vagues antérieures.

Le poids longitudinal ajusté pour les valeurs influentes est donc défini comme étant :

$$w_{infl} = w_{ajust} * \beta_{infl.min}$$

où w_{infl} = poids longitudinal après l'ajustement pour les valeurs influentes

w_{ajust} = poids longitudinal ajusté pour la non-réponse

$\beta_{infl.min}$ = minimum des facteurs d'ajustement pour les valeurs influentes de la vague courante et des vagues antérieures ($0 < \beta_{infl.min} \leq 1$)

4.5 Calage sur marges longitudinal

En théorie, les poids obtenus suite à l'ajustement pour valeurs influentes pourraient servir à produire les estimations d'enquête. Or, on sait, selon la théorie de l'estimation, que s'il existe des données auxiliaires sur la population cible et que celles-ci sont corrélées avec les caractéristiques à l'étude, elles peuvent servir à obtenir des estimations plus précises.

Il existe une autre raison d'utiliser des données auxiliaires lors de l'estimation. Il est souvent important que les estimations d'enquête correspondent à certains totaux connus de la population. De nombreuses enquêtes ajustent leurs estimations par groupe d'âge et selon le sexe de façon à ce qu'elles soient conformes aux estimations projetées basées sur le recensement le plus récent. Comme ces projections démographiques sont révisées suite à la tenue d'un nouveau recensement, l'EDTR procède donc à une révision historique des poids d'enquête environ tous les cinq ans. Il est à noter que l'information auxiliaire utilisée peut aussi provenir de données administratives.

Cette étape de la pondération a subit de nombreux changements depuis les débuts de l'EDTR. Ainsi, dans les premières années de l'EDTR, une simple post-stratification servait à ajuster les poids à des totaux connus de la population. Seuls des groupes d'âge selon le sexe au niveau provincial étaient utilisés (Lévesque et Franklin, 2000). Par la suite, au cours d'un vaste projet d'harmonisation du calage pour les enquêtes sur les dépenses et le revenu et suite à la disponibilité de nouveaux totaux de contrôle, l'EDTR a subit deux importantes révisions historiques en 2002 et 2005. Les nouvelles stratégies de calage ont été développées principalement pour la pondération transversale mais ont aussi été adaptées à la pondération longitudinale. Pour plus de détails sur les raisons justifiant le choix de la stratégie de calage transversal actuelle, voir la section 5.7 sur le calage sur marges transversal.

Le choix des totaux de contrôle qui sont utilisés dans le calage dépend de leur disponibilité et de la qualité de ces derniers. Les données administratives sont sujettes aux erreurs et il faut s'assurer que l'ajustement soit fait sur des données de qualité raisonnable. Ainsi, même si les totaux de contrôle utilisés comportent aussi des erreurs, celles-ci doivent être moins importantes que l'erreur faite en n'ajustant pas les estimations d'enquête. Une autre contrainte quant au choix des totaux de contrôle pouvant être utilisés est que les variables correspondantes dans l'enquête doivent être disponibles pour tous les répondants.

Le tableau 4.3 présente l'historique des totaux de contrôle utilisés pour la pondération de l'EDTR depuis les débuts de l'enquête jusqu'à la dernière révision historique qui a eu lieu en 2005.

Tableau 4.3 Historique des totaux de contrôle utilisés lors du calage

Totaux de contrôle	Avant révision[1] (post-stratification)	Révision historique de 2002 (AR 2000)	Révision historique de 2005 (AR 2003)
Groupes d'âge et de sexe	0 à 6 7à 15 16 à 18 19 à 24 25 à 34 35 à 44 45 à 54 55 à 59 60 à 64 65 à 69 70 et plus	0 à 6 7 à 15 16 à 17 18 à 24 25 à 34 35 à 44 45 à 54 55 à 59 60 à 64 65 à 69 70 et plus	0 à 6 (les deux sexes) 7 à 17 18 à 24 25 à 34 35 à 44 45 à 54 55 à 64 65 et plus
Nombre de ménages selon la taille[2]	...	Taille = 1, Taille = 2	Taille = 1, Taille = 2
Nombre d'entités économiques selon la taille[3]	...	Taille = 1, Taille = 2	Taille = 1, Taille = 2
Nombre de salariés par classes de salaires	...	...	7 classes définies selon les 10e, 25e, 50e, 65e, 75e, 95e ou 98e ou 99e centiles
Période affectée par la révision	...	1980 à 2000	1990 à 2003

... n'ayant pas lieu de figurer

1. Pour le panel 1, d'autres groupes d'âge ont été utilisés dans la post-stratification (Lévesque et Franklin, 2000).
2. Un ménage est constitué d'une personne ou d'un groupe de personnes occupant un même logement.
3. Une entité économique (appelé aussi famille économique) est constituée d'un groupe de deux personnes ou plus qui vivent dans le même logement et qui sont apparentées par le sang, par alliance, par union libre ou par adoption. Dans ce document, une entité économique de taille un (appelé aussi personne seule) est une personne vivant seule ou avec une ou plusieurs autres personnes non apparentées par le sang, par alliance, par union libre ou par adoption.

4.5.1 Théorie du calage

L'EDTR utilise un estimateur par régression généralisée (GREG) dans l'étape du calage (Särndal et coll. 1992). Cet estimateur permet d'ajuster les estimations d'enquête à des totaux connus de la population, appelés totaux de contrôle, tout en minimisant la distance entre les poids avant et après calage et en tenant compte de certaines contraintes sur les poids finaux. C'est une méthode d'estimation assistée par modèle qui permet de caler selon plusieurs variables simultanément. La restriction imposée sur les poids finaux de l'EDTR est que ceux-ci doivent être égaux ou supérieurs à 1 à la fin du calage.

Soit *Y*, la variable d'intérêt et $\underline{y}$, le vecteur *1* x *n* des valeurs observées dans l'échantillon. La notation utilisée est similaire à celle utilisée par Bankier (2002). Soit $\underline{\hat{x}}$ la matrice *p* x *n* des valeurs observées dans l'échantillon pour les *p* variables auxiliaires (pour lesquelles on possède les totaux de contrôle). L'estimateur prend la forme

$$\hat{Y} = \underline{y}\, \mathrm{diag}(\underline{w})\, \underline{g}$$

où $\underline{w}$ est un vecteur *n* x *1* des poids avant calage, $\mathrm{diag}(\underline{w})$ est une matrice diagonale de dimension *n* x n avec les poids avant calage $\underline{w}$ sur la diagonale et finalement $\underline{g}$, qui représente le vecteur des *g*-poids de dimension *n* x *1*. Les *g*-poids sont calculés de sorte que la fonction de perte *L* soit minimisée tout en respectant la contrainte

$$\underline{x}\, \mathrm{diag}(\underline{w})\, \underline{g} = \underline{X}$$

où $\underline{X}$ est le vecteur de dimension *p* x *1* des *p* totaux de contrôle. La fonction de perte utilisée dans l'EDTR est

$$L = (\underline{g} - \underline{1}_n)' \; \underline{1}_n \, (\underline{g} - \underline{1}_n)$$

où $\underline{I}_n$ est la matrice-identité. Les *g*-poids résultants sont donnés par

$$\underline{g} = \underline{1}_n + \underline{\hat{x}}' (\underline{\hat{x}}\underline{\hat{x}}')^{-1} (\underline{X} - \underline{\hat{x}}' \underline{1}_n).$$

4.5.2 Application du calage sur marges pour les poids longitudinaux

Lors de la pondération longitudinale, l'étape du calage est effectuée indépendamment pour chaque panel puisque les estimations d'enquête doivent être ajustées de façon à représenter la population cible au moment de la sélection du panel, c'est-à-dire au 31 décembre 1998 et 2001 pour les panels 3 et 4 respectivement.

Le classement des individus dans les différentes classes de calage se fait en utilisant les caractéristiques des individus longitudinaux au moment de la sélection du panel, soit à la vague 0. Ainsi, l'information disponible pour tous les répondants provient de l'enquête sur la population active (EPA). Par exemple, la province utilisée lors du calage longitudinal représente la province habitée par l'individu au moment de la sélection de l'échantillon d'un panel donné, et le groupe d'âge auquel un individu appartient est déterminé en fonction de son âge toujours au moment de la sélection de l'échantillon du panel concerné.

Tel que mentionné précédemment, la stratégie de calage actuelle a d'abord été développée en fonction de l'aspect transversal de l'enquête. Le tableau 4.3 présente les totaux de contrôle sélectionnés lors de la mise en œuvre de la plus récente stratégie de calage. On y trouve des groupes d'âge selon le sexe, des effectifs selon la taille du ménage et de l'entité économique, ainsi que des effectifs de salariés selon la classe de salaire.

L'utilisation des mêmes totaux de contrôle que ceux sélectionnés pour l'aspect transversal n'a pu être faite de façon directe pour l'aspect longitudinal. En effet, il n'est pas possible de classer les individus longitudinaux dans les différentes classes de salaires et traitements à la vague initiale puisque l'information n'est pas disponible dans l'EPA. Le tableau 4.4 illustre les totaux et variables disponibles aux différentes vagues pour le calage longitudinal :

Tableau 4.4 Totaux de contrôle et variables d'enquête disponibles pour le calage longitudinal

Disponibilité pour calage		Vague 0	Vague 1	Vague X
Variables d'enquête	Démographique	✔	✔	✔
	Salaires et traitements	...	✔	✔
Totaux de contrôle	Démographique	✔	✔	✔
	Salaires et traitements	✔	✔	✔

⇧

Population cible

Pour contourner le problème causé par la non-disponibilité de la variable d'enquête sur les salaires et traitements à la vague 0, plusieurs options ont été proposées et évaluées. La solution retenue a été de faire un calage en deux étapes en utilisant l'information disponible provenant de la première vague. La première étape du calage consiste à ajuster les estimations d'enquête correspondantes à la vague 1 aux totaux de contrôle pour cette même période. Pour cette première étape, tous les totaux de contrôle sont utilisés. Par la suite, la deuxième étape consiste à réajuster les estimations d'enquête de façon à ce qu'elles représentent la population cible, soit celle de la population à la vague 0. Seuls les totaux de contrôle démographiques sont utilisés puisqu'ils sont disponibles dans l'EPA. Comme les totaux démographiques varient peu entre deux années consécutives, compléter le second calage sans les classes de salaires a peu d'impact sur la distribution salariale résultante. Le deuxième calage permet donc d'ajuster les poids de façon à représenter la population initiale tout en gardant une distribution salariale acceptable.

Tableau 4.5 Exemple des totaux de contrôle utilisés pour le calage du panel 4

Totaux de contrôle			
1er calage Vague 1 Période de référence: 31 décembre 2002	Groupe d'âge selon le sexe	1 à 7 (les deux sexes) 8 à 18 19 à 25 26 à 35	36 à 45 46 à 55 56 à 65 66 et plus
	Effectifs selon la taille du ménage	Taille=1 et taille=2	
	Effectifs selon la taille de l'entité économique	Taille=1 et taille=2	
	Classes de salaires et traitements	7 classes définies selon les 10^e, 25^e, 50^e, 65^e, 75^e, 95^e ou 98^e ou 99^e centiles	
Deuxième calage Vague 0 Période de référence : 31 décembre 2001	Groupe d'âge selon le sexe	0 à 6 (les deux sexes) 7 à 17 18 à 24 25 à 34	35 à 44 45 à 4 55 à 64 65 et plus
	Effectifs selon la taille du ménage	Taille=1 et taille=2	
	Effectifs selon la taille de l'entité économique	Taille=1 et taille=2	

En utilisant les totaux de contrôle présentés dans le tableau 4.5 et en suivant la théorie présentée à la section 4.5.1, le vecteur des poids longitudinaux obtenus à la suite du premier calage est défini comme étant

$$\underline{w}_{cal.1} = \text{diag}(\underline{w}_{infl})\underline{g}_1$$

où $\text{diag}(\underline{w}_{infl})$ est une matrice diagonale de dimension *n x n* avec les poids avant calage, c'est-à-dire les poids après ajustement des valeurs influentes, $\underline{w}_{infl}$, sur la diagonale, et $\underline{g}_1$ représente le vecteur des *g*-poids de dimension *n* x *1* obtenus suite au premier calage.

Le vecteur des poids longitudinaux obtenu à la suite du deuxième calage est donné par

$$\underline{w}_{cal,2} = \text{diag}(\underline{w}_{cal,1})\underline{g}_2$$

où $\text{diag}(\underline{w}_{cal,1})$ est une matrice diagonale de dimension *n* x *n* avec les poids résultant du premier calage, $\underline{w}_{cal,1}$, sur la diagonale, et $\underline{g}_2$ représente le vecteur des *g*-poids de dimension *n* x *1* obtenus du deuxième calage.

L'ajustement moyen apporté aux poids longitudinaux ajustés pour la non-réponse et les valeurs influentes, suite à l'étape du calage est de 1,43 pour la cinquième vague du panel 3 et de 1,4 pour la deuxième vague du quatrième panel.

L'ensemble de poids résultant du calage en deux étapes constitue l'ensemble des poids finaux longitudinaux de l'EDTR[7]. Le tableau 4.6 donne les médianes des poids longitudinaux finaux par province pour l'année de référence 2003. La province donnée dans le tableau est la province habitée par l'individu au 31 décembre de l'année de référence, c'est-à-dire 2003. Les individus qui étaient décédés, institutionnalisés ou hors de la population des dix provinces canadiennes au 31 décembre 2003 sont inclus dans une catégorie à part.

Tableau 4.6 Médianes des poids longitudinaux finaux, pour la cinquième vague du panel 3 et la deuxième vague du panel 4, année de référence 2003

Province	Longitudinaux finaux	
	Panel 3	Panel 4
	poids	
Terre Neuve et Labrador	312	331
Île-du-Prince-Édouard	131	134
Nouvelle-Écosse	364	354
Nouveau-Brunswick	331	351
Québec	794	727
Ontario	833	781
Manitoba	425	448
Saskatchewan	358	354
Alberta	1004	1065
Colombie-Britannique	1283	1129
Décédés, institutionnalisés ou hors des 10 provinces canadiennes	601	680

7. Auparavant, l'EDTR produisait aussi un ensemble de poids longitudinaux externes auxquels on ajoutait du bruit de façon aléatoire, et ce pour usage externe (Lévesque et Franklin, 2000). Ceux-ci ne sont plus produits.

5 Pondération transversale

La présente section donne une description détaillée des étapes nécessaires pour créer un ensemble de poids représentatif de la population des dix provinces canadiennes au 31 décembre d'une année de référence donnée. Tous les individus longitudinaux qui sont dans le champ de l'enquête transversalement à la période de référence et leurs cohabitants sont considérés lors de la production de cet ensemble de poids. La méthodologie présentée dans cette section est celle utilisée pour l'année de référence 2003.

Deux types de poids transversaux sont produits par l'EDTR : le poids de type intégré et le poids de type individu. La différence entre ces deux types de poids est que le poids de type individu peut être différent pour chaque personne à l'intérieur d'un même ménage tandis que le poids intégré est identique pour tous les membres du ménage. Le poids de type intégré est adéquat pour faire des analyses au niveau du ménage ou de la famille, mais peut aussi être utilisé pour des analyses au niveau des individus. Quant au poids de type individu, il permet seulement de faire des analyses au niveau des individus[8].

Le premier ensemble de poids produit, le poids de type intégré ou *poids intégré*, est le poids le plus couramment utilisé pour le calcul des estimations transversales incluses dans les publications de l'EDTR. Le deuxième ensemble de poids produit est le *poids travail*. Ce dernier poids est un poids de type individu pour lequel la définition de répondant utilisée a été quelque peu modifiée par rapport à celle utilisée pour le poids intégré. Tel que le nom du poids l'indique, le poids travail ne considère que la réponse à l'entrevue sur le travail, au lieu d'utiliser la réponse aux deux entrevues (entrevue sur le travail et entrevue sur le revenu) pour identifier un répondant. Par exemple, un ménage non-répondant à l'entrevue sur le travail (i.e. ménage pour lequel tous ses membres sont non-répondants à cette entrevue) mais répondant à l'entrevue sur le revenu (par exemple, à cause d'une permission donnée préalablement pour utiliser leurs données fiscales) se verra assigner un poids travail nul lors de la pondération. L'utilisation de ce poids, au lieu d'un poids intégré, permet d'avoir une meilleure estimation des caractéristiques sur le travail.

La pondération transversale s'effectue selon les sept étapes suivantes : la détermination des individus éligibles à la pondération transversale (section 5.1), l'ajustement pour la non-réponse (section 5.2), l'ajustement pour la migration interprovinciale (section 5.3), l'application des facteurs d'allocation des panels (section 5.4), le partage de poids (section 5.5), l'ajustement pour les valeurs influentes (section 5.6) et finalement le calage sur marges qui constitue la dernière étape menant à la création des poids transversaux finaux (section 5.7).

5.1 Détermination des individus éligibles à la pondération transversale

L'échantillon transversal pour une année donnée est représentatif de la population des dix provinces canadiennes au 31 décembre de l'année de référence (à l'exception des ménages constitués exclusivement d'immigrants). Cet échantillon est composé des individus longitudinaux faisant toujours partie de la population des dix provinces canadiennes au 31 décembre de l'année de référence, ainsi que de leurs cohabitants. On dit qu'une personne est *éligible transversalement* si elle fait partie de cet échantillon.

Pour déterminer quels individus auront un poids transversal non nul, il faut d'abord identifier ceux qui sont éligibles transversalement. Tous les individus longitudinaux qui ne font plus partie de la population cible au 31 décembre de l'année de référence (comme les personnes ayant déménagé en dehors des dix provinces canadiennes, les personnes décédées ou en institution) reçoivent un poids transversal nul, même s'ils sont répondants (ou considérés comme répondants). Pour les individus longitudinaux qui font partie de la population cible au 31 décembre de l'année de référence, seuls ceux qui sont dans un ménage répondant ont un poids transversal non nul. Pour la production des poids intégrés, la notion de ménage répondant est la suivante : un ménage répondant est un ménage dans lequel il y a au moins une personne qui a répondu à au

8. Auparavant, l'EDTR produisait des ensembles de poids à usage interne seulement et d'autres à usage externe. Cependant, les poids d'usage externe ne sont plus créés et seuls les poids internes sont maintenant utilisés (Lévesque et Franklin, 2000).

moins une des deux entrevues (entrevue sur le travail, entrevue sur le revenu). Il a été déterminé en outre que la composition d'un ménage ne peut être manquante pour plus d'un an. Ainsi, si un ménage est non répondant à l'entrevue sur le travail pour deux années consécutives, alors celui-ci sera considéré comme non-répondant. Cette dernière condition a été ajoutée lors de la révision historique qui a eu lieu en 2003 pour éviter d'avoir un grand nombre de ménages répondants pour lesquels il n'y a aucune information récente sur la composition du ménage. Ceci élimine surtout les cas pour lesquels il y a une permission d'apparier aux fichiers d'impôt mais aucune autre information disponible. Les personnes non-répondantes dans un ménage répondant auront de l'information imputée. Pour le poids travail, un ménage répondant est un ménage dans lequel il y a au moins une personne qui a répondu à l'entrevue sur le travail.

5.2 Ajustement pour la non-réponse

Comme mentionné précédemment, l'échantillon transversal pour une année donnée est représentatif de la population des dix provinces canadiennes au 31 décembre de l'année de référence. Cependant, comme dans toute enquête, l'EDTR est sujette à la non-réponse (on parle ici de non-réponse totale) et la représentativité de l'échantillon transversal en est alors affectée. Pour réduire l'effet du biais causé par la non-réponse, on ajuste les poids des répondants pour compenser pour les non-répondants.

Du point de vue transversal, on détient trop peu d'information sur les cohabitants non-répondants pour permettre un ajustement pour la non-réponse à ce niveau. En effet, pour plusieurs cohabitants, les seules variables disponibles sont la province, l'âge et le sexe, variables qui seront de toute façon utilisées à l'étape du calage. Considérant cela, deux méthodes d'ajustement pour la non-réponse au niveau transversal ont été proposées dans le passé (Latouche et Michaud, 1997). La première solution consistait tout d'abord à faire un ajustement pour la non-réponse pour les individus longitudinaux seulement, et à ensuite utiliser la méthode du partage de poids pour attribuer un poids aux cohabitants. De cette manière, les poids des cohabitants étaient indirectement ajustés pour la non-réponse. La deuxième alternative consistait à faire un ajustement pour la non-réponse au niveau des ménages en se servant des strates et des groupes de renouvellement des logements initiaux selon le plan de sondage de l'EPA. L'ajustement au niveau des ménages tendait toutefois à sous-estimer légèrement les taux de réponse.

La première méthode décrite ci-haut est celle qui a été adoptée par l'enquête. L'étape d'ajustement transversal pour la non-réponse ne relève donc pas de la pondération transversale, mais plutôt de la pondération longitudinale. La modélisation et l'ajustement pour la non-réponse longitudinale ont été décrits à la section 4.3. Rappelons tout de même que, pour la pondération longitudinale, chaque panel est traité séparément. Le poids transversal ajusté pour la non-réponse longitudinale est donc égal au poids w_{ajust} dérivé à la section 4.3. Dénotons par $w_{2003,ajust}$, le poids transversal ajusté pour la non-réponse longitudinale pour l'année de référence 2003. Ce poids est utilisé comme poids initial pour la production de tous les poids transversaux dérivés (poids intégré et poids travail).

Tel que mentionné précédemment, le groupe des non-répondants transversaux a peu d'importance ici car l'ajustement de non-réponse se fait lors de la pondération longitudinale. Ainsi seuls les ménages longitudinaux considérés répondants auront un poids non nul à ce stade-ci et les cohabitants faisant partie de ces ménages seront considérés dans la suite du processus de pondération transversale.

5.3 Ajustement pour la migration interprovinciale

L'étape de l'ajustement pour la migration interprovinciale est une autre étape directement engendrée par la nature longitudinale de l'EDTR. À travers les années, il arrive que certaines personnes échantillonnées déménagent d'une province à une autre. Étant donné que la probabilité de sélection peut varier beaucoup selon la géographie, un grand écart est observé, à l'échelle canadienne, entre les poids que l'on retrouve dans les différentes provinces. Par conséquent, si une personne avec une probabilité de sélection assez faible (donc un poids assez grand) déménage dans un endroit où la probabilité de sélection des gens est beaucoup plus forte (donc des poids beaucoup plus petits), il y a potentiellement un problème de précision des

estimations provinciales. Du point de vue analytique, les résultats d'analyse peuvent se trouver faussés à cause de certains cas de migration interprovinciale.

Pour expliquer le problème, prenons un exemple simple. Supposons qu'un chirurgien cardiaque faisant partie de l'échantillon déménage de Toronto à Charlottetown. Cette situation causerait plusieurs problèmes. D'abord, son poids qui s'établirait aux environs de 1 100 serait beaucoup plus grand que tous les poids que l'on retrouve à l'Île-du-Prince-Édouard et qui sont en moyenne de 65. Même si à lui seul, il ne représenterait qu'environ 0.7% de la population de la province, cette valeur de poids combinée avec un haut revenu viendrait gonfler artificiellement le revenu total des individus à l'Île-du-Prince-Édouard. Notons qu'en pratique, il n'est aussi pas réaliste que 1 100 chirurgiens cardiaques décident tous de déménager de Toronto à Charlottetown. Ceci n'est donc pas acceptable d'un point de vue analytique.

Plusieurs options sont envisageables pour atténuer le problème de poids aberrants causés par la migration interprovinciale. Une solution est de post-stratifier l'échantillon selon des classes formées par le croisement des provinces d'origines et de destinations des individus. Ceci nécessite cependant les totaux de contrôles des migrations interprovinciales et est aussi relativement difficile à implanter. Ces poids extrêmes peuvent aussi être remplacés par le poids moyen, la médiane ou un des quantiles de la distribution des poids des personnes de la province où les gens déménagent, ou toute autre possibilité basée sur cette distribution. Ces alternatives sont arbitraires et pourraient toutes être utilisées. Il s'agit en fait de trouver un compromis entre le fait d'attribuer à la personne migrante un poids similaire à la moyenne des poids des personnes de sa province de destination et le fait de ne pas trop modifier son poids d'origine (basé sur la probabilité de sélection). L'objectif ici est tout d'abord d'identifier les personnes qui ont migré dans une autre province et dont le poids contribue trop fortement à l'estimation de leur province de destination, et ensuite d'ajuster ce poids.

L'EDTR a adopté une méthode utilisant le 95^{e} percentile de la distribution des poids. Cette méthode a l'avantage d'être simple, rapide et de ne pas nécessiter d'intervention manuelle. Il s'agit tout d'abord d'identifier toutes les personnes qui ont déménagé dans une province particulière dont leur poids est plus grand que le maximum des poids des personnes originaires de cette province. Si c'est le cas, le poids de ces personnes est ramené au 95^{e} percentile de la distribution des poids des personnes originaires de la province de destination. Ensuite, un nombre aléatoire entre 0.00 et 10.00 est soustrait de ce percentile afin de conserver la "presque" unicité des poids (il est possible d'avoir des poids individus égaux mais cela est très rare).

Lors de la dernière révision historique, des changements ont été apportés à cette étape de la pondération. Auparavant, les ajustements étaient effectués après l'étape de partage des poids et indépendamment pour les poids individus et intégrés (voir Lévesque et Franklin, 2000). La façon de faire les ajustements était différente puisque dans le cas du poids intégré, il fallait se soucier de préserver l'égalité des poids à l'intérieur du ménage.

La nouvelle méthodologie employée est appliquée à la suite de l'ajustement de non-réponse. Ceci permet de ne faire qu'un seul type d'ajustement au niveau individuel et ce, pour tous les ensembles de poids transversaux. En changeant ainsi l'ordre dans lequel cette étape est effectuée, les ajustements pour la migration interprovinciale ne sont effectués que sur les individus longitudinaux sur une base individuelle. Ainsi, seuls les individus longitudinaux ayant migrés dans une autre province et dont le poids ajusté pour la non-réponse dépasse le maximum des poids de la province de destination voient leur poids ajusté pour la migration interprovinciale. Comme les poids de ces individus longitudinaux sont partagés par la suite avec leurs cohabitants, l'ajustement est indirectement appliqué sur les autres membres du ménage.

Les ajustements pour la migration interprovinciale sont effectués et sont redérivés à chaque année de traitement. Le poids transversal ajusté pour la migration interprovinciale pour l'année de référence 2003 est défini comme étant :

$$w_{2003.mig} = w_{2003.ajust} \cdot a_{mig}$$

où $w_{2003,mig}$ = poids transversal ajusté pour la migration interprovinciale

$w_{2003,ajust}$ = poids transversal ajusté pour la non-réponse

a_{mig} = facteur d'ajustement pour la migration interprovinciale.

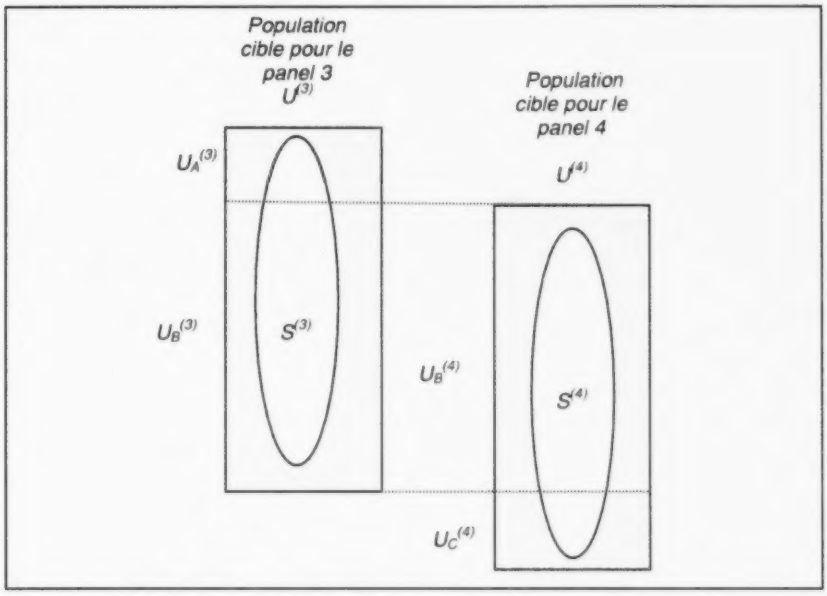
Pour l'année de référence 2003, un ajustement de migration interprovinciale s'est avéré nécessaire pour 94 personnes longitudinales du panel 3 et 34 personnes longitudinales du panel 4. Comme l'année de référence 2003 correspond à la cinquième vague du panel 3 et seulement à la deuxième vague du panel 4, ceci explique pourquoi il y a un plus grand nombre de personnes longitudinales ajustées pour la migration interprovinciale dans le panel le plus ancien.

5.4 Application des facteurs d'allocation des panels

À partir de l'année de référence 1996, l'EDTR a commencé à avoir deux panels qui se chevauchent. En théorie les deux panels peuvent être utilisés indépendamment l'un de l'autre pour produire des estimations transversales, mais en pratique il est plus avantageux de combiner les échantillons transversaux des deux panels afin de produire les estimations transversales. La taille d'échantillon est alors doublée, ce qui réduit la variabilité des estimations et offre la possibilité d'utiliser plus de totaux de contrôle lors de l'étape du calage sur marges.

Comme les deux échantillons longitudinaux sont représentatifs de la population des dix provinces canadiennes à des périodes de référence différentes (31 décembre 1998 pour le troisième panel et 31 décembre 2001 pour le quatrième panel), il faut trouver un moyen permettant de combiner les deux échantillons pour que ces derniers représentent ensemble la population cible d'une année référence donnée. Le problème qui se pose dans la combinaison des échantillons de deux panels consécutifs est le temps écoulé entre la sélection des deux échantillons. La figure 5.1 présente la dynamique de la population des dix provinces canadiennes entre la sélection des panels 3 et 4 de l'EDTR.

Figure 5.1 Dynamique de la population des dix provinces canadiennes entre la sélection des panels 3 et 4



En janvier 1999, on a tiré un échantillon $S^{(3)}$ de la population des dix provinces canadiennes $U^{(3)}$. Cet échantillon est l'échantillon longitudinal associé au troisième panel. En janvier 2002, un échantillon $S^{(4)}$ a été sélectionné de la population des dix provinces canadiennes $U^{(4)}$ pour former l'échantillon longitudinal du quatrième panel. La majorité des individus faisant partie de la population en janvier 1999 sont aussi présents dans la population en janvier 2002 ($U_B^{(3)} = U_B^{(4)}$). Les différences entre les populations $U^{(3)}$ et $U^{(4)}$ sont causées par les personnes qui sont sorties de la population cible entre janvier 1999 et janvier 2002 (personnes décédées, institutionnalisées, personnes ayant déménagé à l'extérieur des provinces, etc.) et par les nouveaux arrivants dans la population après janvier 1999 (immigrants, naissances, etc.). Une partie de l'échantillon du troisième panel comprend des personnes de la population sortante entre janvier 1999 et janvier 2002 ($U_A^{(3)}$); les individus de cette population ne font alors plus partie de la population canadienne au moment de la sélection de l'échantillon du quatrième panel et n'ont ainsi aucune chance d'être choisis. De même, une partie de l'échantillon du quatrième panel est composée de personnes qui sont dans la population des dix provinces canadiennes au moment de la sélection de cet échantillon mais qui n'étaient pas présentes dans la population lors de la sélection du troisième panel ($U_C^{(4)}$); ces personnes n'avaient eu aucune chance d'être choisies à ce moment.

Afin de développer une méthode qui permet de combiner l'information des échantillons transversaux des deux panels, il suffit de considérer les deux panels au moment de l'introduction du quatrième panel. La population que l'on doit considérer est $U^{(4)}$, celle effective lors de l'introduction du quatrième panel. Les échantillons longitudinaux à considérer sont les échantillons originaux choisis, l'échantillon du troisième panel ayant été réduit à cause des individus qui ont quitté la population et des non-répondants. La méthode développée par Merkouris (1999) est celle que l'EDTR utilise pour jumeler ses échantillons qui se chevauchent (dans le cas présent, il s'agit des échantillons des panels 3 et 4). Soit Y, un paramètre de la population. Pour estimer Y, il s'agit d'utiliser l'estimateur combiné suivant (qui donne une importance relative à chacun des panels selon un certain facteur p) :

$$\hat{Y} = p\,\hat{Y}_3 + (1-p)\,\hat{Y}_4$$

où $\hat{Y}$ est l'estimation transversale pour une variable d'intérêt, $\hat{Y}_i$ est l'estimation produite pour cette même variable d'intérêt pour le panel i, et p est le facteur d'allocation des panels. Cet estimateur peut être vu comme une somme pondérée des estimateurs des deux panels où les poids sont les facteurs d'allocation des panels. La valeur optimale du facteur d'allocation des panels a été déterminée de façon à ce que la variance de $\hat{Y}$ soit minimale. Le facteur d'allocation des panels 3 et 4 est défini comme suit :

$$p = \left(\frac{n_1}{d_1}\right)\left(\frac{n_1}{d_1} + \frac{n_2}{d_2}\right)^{-1} = \frac{n_1}{n_1 + n_2\left(\frac{d_1}{d_2}\right)} \qquad (5.1)$$

où d_i désigne l'effet de plan d'échantillonnage pour le panel i et n_i représente le nombre de personnes longitudinales de 16 ans et plus du panel i répondantes au temps t (t étant l'année de référence traitée, ici on parle de l'année de référence 2003). L'effet de plan d'échantillonnage est défini ici comme étant le rapport entre la variance de l'estimation sous-pondérée (i.e. sans étalonnage pour respecter les totaux de contrôle de population) et la variance qui aurait résulté d'un échantillon aléatoire simple de même taille. Pour l'EDTR, le nombre de personnes longitudinales de 16 ans et plus répondantes au temps t correspond au nombre de personnes longitudinales de 16 ans et plus éligibles transversalement qui sont dans un ménage répondant (puisque toutes les personnes dans les ménages répondants sont considérées comme répondantes, voir section 4.2). Même si les cohabitants sont utilisés pour les estimations transversales, seuls les répondants longitudinaux au temps t sont considérés dans le calcul des facteurs d'allocation des panels. Merkouris (1999) et Lavallée (1994) suggèrent de combiner les deux panels avant d'intégrer les cohabitants par la méthode du partage de poids. Omettre les cohabitants dans le calcul des facteurs p contribue à réduire l'importance du panel le plus âgé, et d'ainsi minimiser les impacts potentiels dus au biais d'érosion.

La détermination d'un facteur p optimal nécessite donc l'estimation d'effets de plan d'échantillonnage et de tailles d'échantillons longitudinaux. Comme les effets de plan de l'EDTR ne sont généralement pas disponibles lors du traitement des données et que l'EDTR est un sous-échantillon de l'EPA, il a été décidé d'utiliser les données de l'EPA pour obtenir les effets de plan d'échantillonnage nécessaires pour le calcul des facteurs p. De plus, comme on s'intéresse au rapport entre deux effets de plan, l'EPA est tout aussi adéquat pour obtenir les effets de plan désirés. Pour déterminer les effets de plan nécessaires, les six groupes de renouvellement de l'EPA sont utilisés. Ces effets de plan sont calculés pour la caractéristique *nombre de personnes de 15 ans et plus* à l'échelle provinciale (ce qui implique le calcul de dix facteurs d'allocation des panels, c'est-à-dire un par province).

Pour la combinaison des premier et deuxième panels, les fichiers de l'EPA de janvier 1993 et de janvier 1996 ont été utilisés pour calculer les effets de plan (Lévesque et Franklin, 2000). Ces deux panels provenaient de plans de sondage différents puisque l'EPA a fait l'objet d'un remaniement post-censitaire entre la sélection du premier et du deuxième panel (ce dernier s'est terminé à la fin de 1994). En ce qui concerne les panels 3 et 4, il a été décidé de ne pas utiliser les effets de plan dans le calcul des facteurs d'allocation puisque les deux panels ont été sélectionnés selon le même plan de sondage et que les effets de plan de l'EPA étaient donc très similaires. Le rapport des effets de plan était donc de 1. De plus, le vrai ratio a tendance à favoriser l'ancien panel.

Les tailles d'échantillons voulues ont été calculées à partir des échantillons longitudinaux des deux panels pour l'année de référence traitée. Comme mentionné précédemment, les tailles d'échantillons désirées sont le nombre de personnes longitudinales de 16 ans et plus répondantes au temps t. On les obtient en considérant toutes les personnes longitudinales dans les ménages répondants qui ont 16 ans et plus et qui sont éligibles transversalement.

Connaissant les effets de plan et les tailles d'échantillons, il est facile de calculer les facteurs d'allocation des panels d'après l'équation (5.1). Ces facteurs doivent être calculés pour chaque année traitée puisque la composition des échantillons longitudinaux change à chaque année. Pour des raisons pratiques, il a été

décidé de calculer les effets de plan à chaque introduction d'un nouveau panel, et d'utiliser ces effets de plan dans le calcul des facteurs d'allocation des panels pour les trois années de chevauchement (Latouche et coll., 2000). Le tableau 5.1 qui suit donne les facteurs d'allocation des panels calculés pour l'année de référence 2003.

Tableau 5.1 Facteurs d'allocation des panels pour l'année de référence 2003

Province	d_3/d_4 [1]	n_3	n_4	P Panel 3	$(1-P)$ Panel 4
			pourcentage		
Terre-Neuve et Labrador	1	1214	1139	0.5159	0.4841
Ile-du-Prince-Édouard	1	751	820	0.4780	0.5220
Nouvelle-Écosse	1	1793	1822	0.4960	0.5040
Nouveau-Brunswick	1	1620	1571	0.5077	0.4923
Québec	1	4972	5386	0.4800	0.5200
Ontario	1	7570	8288	0.4774	0.5226
Manitoba	1	1990	1961	0.5037	0.4963
Saskatchewan	1	1924	1923	0.5001	0.4999
Alberta	1	2197	2246	0.4945	0.5055
Colombie-Britannique	1	2252	2539	0.4700	0.5300

1. Pour la combinaison des panels 3 et 4, un rapport entre les effets de plan de 1 a été appliqué puisqu'il y avait peu de différences entre les effets de plan de l'EPA pour ces deux périodes.

d_i = effet de plan d'échantillonnage pour le panel *i*.

n_i = nombre de personnes longitudinales de 16 ans et plus du panel *i* répondantes pour l'année de référence 2003.

p, $1 - p$ = facteurs d'allocation des panels.

D'après ce tableau, on constate que les facteurs d'allocation des panels sont proches de 0.5, ce qui fait qu'on accorde une importance similaire aux deux panels, contrairement à ce qui a été fait lors du calcul des facteurs d'allocation entre les panels 1 et 2 (voir Lévesque et Franklin, 2000).

Les facteurs *p* sont appliqués à toutes les personnes longitudinales du troisième panel qui sont dans un ménage répondant au temps *t* (dans ce cas-ci, au 31 décembre 2003) et qui faisait partie de la population cible au moment de la sélection du quatrième panel. Ainsi, le facteur *p* n'est pas appliqué pour les personnes du panel 3 qui étaient institutionnalisées ou qui vivaient hors des 10 provinces canadiennes au moment de la sélection du panel 4 puisque ces dernières n'avaient aucune chance d'être sélectionnées dans l'échantillon du quatrième panel. Avant la dernière révision historique, le facteur d'allocation *p* était appliqué à tous les individus du panel le plus ancien (Lévesque et Franklin, 2000). Cette modification a été appliquée de façon rétroactive pour toutes les années de l'EDTR.

Les facteurs *1-p* devraient normalement être appliqués à toutes les personnes longitudinales du quatrième panel qui sont dans un ménage répondant au temps *t*, à l'exception des personnes en dehors de la population cible au moment de l'introduction du troisième panel (en janvier 1999) et présents dans la population cible lors de l'introduction du quatrième panel (en janvier 2002). Pour des raisons pratiques, seuls les nouveaux-nés et les immigrants[9] absents lors de l'introduction du troisième panel et présents lors de l'introduction du quatrième panel sont exclus lors de l'application des facteurs *1-p*. Les personnes exclues conserveront donc leur poids

9. Les questions sur l'immigration ne sont posées qu'aux personnes âgées de 16 ans et plus lors de l'entrevue de janvier. Ainsi, le statut d'immigrant et l'année d'immigration sont imputés à partir de l'information du plus proche parent pour pouvoir identifier les enfants nés avant la sélection du panel le plus ancien et qui ont immigré dans l'une des dix provinces canadiennes durant la période incluse entre la sélection des deux panels. Lorsque les informations sont obtenus pour un enfant (c'est-à-dire lorsqu'il a 16 ans), celles-ci remplace dorénavant les valeurs ayant été imputées.

ajusté pour la migration interprovinciale. Ainsi le poids transversal ajusté pour la combinaison des deux panels est le suivant :

$$w_{2003,p} = \begin{cases} p\, w_{2003,mig} & \text{Pour les personnes du troisième panel non exclues pour l'application des facteurs d'allocation des panels,} \\ (1-p)\, w_{20} & \text{Pour les personnes du quatrième panel non exclues pour l'application des facteurs d'allocation des panels,} \\ w_{2003,mig} & \text{Pour les personnes des troisième et quatrième panels exclues pour l'application des facteurs d'allocation des panels} \end{cases}$$

où $w_{2003,p}$ = poids transversal ajusté pour la combinaison des deux panels.

$w_{2003,mig}$ = poids transversal ajusté pour la migration interprovinciale.

p = facteur d'allocation des panels.

Pour le poids intégré et le poids travail, le poids ajusté pour la combinaison des panels se calcule de la même façon. Cependant, la notion de ménage répondant appliquée dans le cas du poids travail n'est pas la même et celle-ci est présentée à la section 5.1. Il est à noter que les facteurs d'allocation des panels sont appliqués aux poids après l'étape d'ajustement pour la non-réponse et pour la migration interprovinciale, mais avant l'étape du partage de poids pour que ces facteurs soient aussi inclus dans le poids des cohabitants.

5.5 Partage des poids

L'étape du partage des poids est essentielle et unique à la pondération transversale. Elle est engendrée par la nature longitudinale de l'EDTR. Dans une enquête longitudinale, l'échantillon évolue à travers le temps. De nouvelles personnes viennent habiter avec les individus déjà sélectionnés, des enfants naissent dans certains ménages échantillonnés. Bref, d'autres personnes qui n'ont pas été échantillonnées initialement sont désormais interviewées. Puisqu'on ne connaît pas la probabilité de sélection des personnes qui se sont jointes à l'enquête (les cohabitants), il faut trouver une façon de leur assigner un poids. Cela est fait en utilisant la méthode du partage des poids. Les prochains paragraphes expliquent brièvement la méthodologie du partage des poids. Pour plus de détails sur cette méthode, voir Lavallée (1995) et Lavallée (2002).

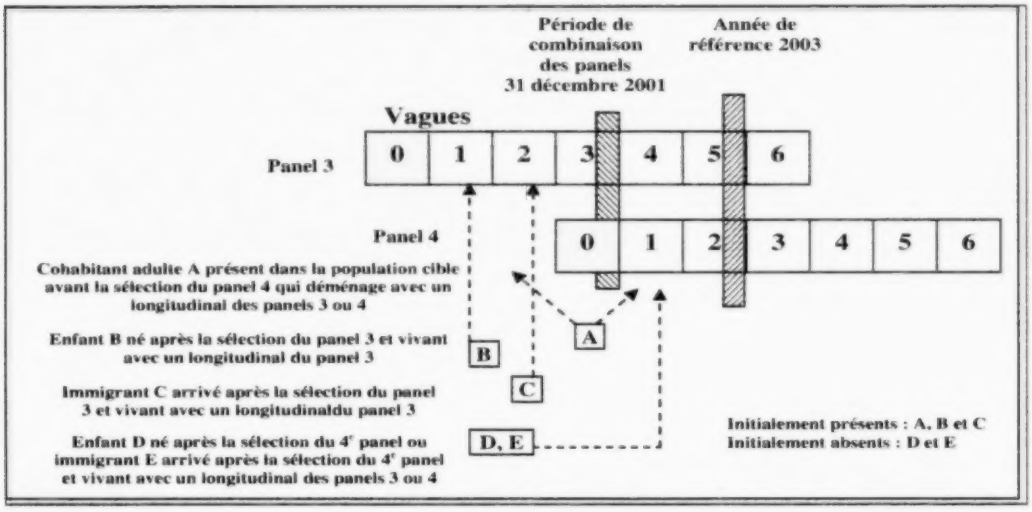
Pour effectuer le partage des poids, il faut d'abord déterminer à quel niveau dans la hiérarchie du plan de sondage on veut effectuer ce partage : au niveau de la province, de la strate, de l'unité primaire d'échantillonnage ou de l'unité finale d'échantillonnage. L'EDTR a choisi d'effectuer le partage des poids au niveau de l'unité finale d'échantillonnage, c'est-à-dire le ménage. Une fois cela décidé, il s'agit de déterminer le statut de chaque personne dans l'échantillon transversal, selon que la personne est un individu longitudinal, un cohabitant initialement présent ou un cohabitant initialement absent.

À partir de l'étape du partage de poids, on ne fait plus de distinction entre les deux panels et on traite tous les individus comme faisant partie d'un même échantillon, tout en conservant les notions d'individu longitudinal et de cohabitant. On pourrait penser qu'en choisissant d'effectuer le partage de poids au niveau du ménage, traiter les deux échantillons transversaux de façon individuelle ou de façon combinée reviendrait au même. Or, il faut penser qu'en présence d'une enquête à panels multiples se chevauchant, il est possible d'observer des ménages pour lesquels des personnes issues du panel le plus ancien habitent avec des personnes issues du panel le plus récent. Lors de l'introduction du quatrième panel, tout ménage sélectionné pour faire partie de l'échantillon longitudinal comportant au moins une personne longitudinale du troisième panel a été exclu de l'échantillon longitudinal du quatrième panel afin de ne pas enquêter le ménage pour une autre période de six ans. Cependant, étant donné la nature longitudinale et dynamique de l'enquête, il se peut qu'au cours des six années, des ménages se séparent et de nouveaux ménages se forment, rendant ainsi possible

le fait d'observer un ménage composé de personnes provenant des deux panels. En considérant les échantillons transversaux séparément, deux ensembles de poids distincts seraient produits pour ces ménages; ce qui causerait des problèmes notamment de cohérence dans les estimations. Il est alors préférable de considérer les deux panels ensemble.

Un cohabitant est initialement présent s'il faisait partie de la population cible au moment de la sélection du dernier (quatrième) panel, soit en janvier 2002 (cohabitant adulte A, enfant B et immigrant C dans la Figure 5.2). À l'inverse, un cohabitant est initialement absent s'il ne faisait pas partie de la population cible au moment de la sélection du quatrième panel (personne D et E dans la Figure 5.2). Étant donné qu'on traite ici le chevauchement de deux panels, les notions de cohabitant initialement absent et initialement présent doivent tenir compte du fait qu'on ne traite plus qu'un seul panel et qu'on doit traiter les deux panels ensemble. On doit alors considérer les deux panels au moment de la sélection du dernier panel, soit le quatrième panel, en janvier 2002. Il est important de mentionner que tous les cohabitants de l'enquête ont un poids longitudinal de zéro étant donné qu'ils ne faisaient pas partie de l'échantillon longitudinal. Par contre, ils peuvent avoir un poids transversal différent de zéro qui leur est assigné par la méthode du partage des poids.

Figure 5.2 Définition des cohabitants initialement absents et présents pour l'année de référence 2003



À l'étape du partage des poids, deux types de poids sont produits pour chaque individu : le poids intégré (poids identique pour tous les membres d'un même ménage) et le poids travail qui est un poids individu (qui peut être différent pour chaque personne à l'intérieur d'un ménage donné). La façon de calculer les poids partagés dépend de la présence de cohabitants dans le ménage ainsi que du type de poids, c'est-à-dire de type intégré ou individu.

5.5.1 Partage des poids pour le poids intégré

Dans le cas du partage des poids pour le poids intégré, les poids résultant du partage doivent être égaux pour tous les membres du ménage, et ce, pour tous les ménages. Ceci est nécessaire dû au fait que le poids avant l'étape du calage doit être identique pour tous les gens d'un ménage afin de pouvoir produire un poids intégré (Lemaître et Dufour, 1987). Or, les poids ajustés pour la combinaison des panels ne sont pas égaux pour tous les membres d'un ménage puisque les ajustements précédents sont principalement fait au niveau individuel.

Pour un ménage *h* donné, le poids transversal partagé intégré est défini comme étant :

$$w_{2003,\,part} = \frac{\sum_{h} w_{2003,\,p}}{n_{L,\,h} + n_{IP,\,h}}$$

où $w_{2003,\,part}$ = Poids transversal partagé

$w_{2003,\,p}$ = Poids transversal ajusté pour la combinaison des deux panels

$n_{L,\,h}$ = Nombre de personnes longitudinales dans le ménage *h*

$n_{IP,\,h}$ = Nombre de cohabitants initialement présents dans le ménage *h*.

5.5.2 Partage des poids pour le poids travail

Puisque le poids travail est un poids individuel (ou non-intégré), le partage des poids est effectué selon les étapes suivantes. D'abord, il faut identifier les ménages auxquels aucun cohabitant ne s'est greffé depuis la sélection de l'échantillon longitudinal. Pour ces ménages composés exclusivement de personnes longitudinales, il n'est pas nécessaire d'appliquer la méthode du partage des poids. Par conséquent, pour tous les individus de ces ménages, le poids transversal partagé est identique au poids transversal ajusté après l'application des facteurs d'allocation des panels.

Pour les ménages où des cohabitants sont arrivés, il s'agit de faire la somme des poids transversaux ajustés pour le facteur d'allocation des panels $w_{2003,p}$ de tous les individus longitudinaux à l'intérieur du ménage. Puis, il faut établir le nombre d'individus longitudinaux $n_{L,h}$ et le nombre de cohabitants initialement présents $n_{IP,h}$ dans chaque ménage *h*.

Pour les ménages où au moins un des cohabitants est initialement présent, une modification à la méthodologie a été apportée lors de la dernière révision historique afin de préserver l'ajustement de non-réponse des individus longitudinaux, ainsi que le rapport entre les poids des individus longitudinaux du ménage. Pour plus de détails sur l'ancienne méthodologie employée, se référer à Lévesque et Franklin (2000). Le poids partagé est maintenant calculé différemment pour les individus longitudinaux et pour les cohabitants. Pour les cohabitants, le poids transversal partagé est calculé comme étant la somme des poids transversaux $w_{2003,p}$ de tous les individus longitudinaux du ménage divisée par le nombre total d'individus longitudinaux et de cohabitants initialement présents dans le ménage. Pour les individus longitudinaux, le poids transversal partagé est égal au poids avant partage $w_{2003,p}$, multiplié par un facteur équivalent au nombre d'individus longitudinaux divisé par la somme du nombre d'individus longitudinaux et de cohabitants initialement présent.

Pour les ménages où tous les cohabitants sont initialement absents, le poids transversal partagé des individus longitudinaux et des cohabitants est calculé différemment. Pour les individus longitudinaux faisant partie de ces ménages, le poids transversal partagé est égal au poids transversal ajusté après l'application des facteurs d'allocation des panels. Pour les cohabitants, le poids transversal partagé est égal à la somme des poids $w_{2003,p}$ de tous les individus longitudinaux du ménage, divisé par le nombre total d'individus longitudinaux. De cette façon, un poids moyen n'est pas assigné aux individus longitudinaux et l'ajustement pour la non-réponse est préservé. Seuls les individus initialement absents reçoivent le poids moyen des individus longitudinaux du ménage.

Pour un ménage *h* donné, le poids transversal individu partagé (poids individu usuel ou poids travail) est ainsi défini comme étant :

$$w_{2003,part} = \begin{cases} w_{2003,p} \cdot & \text{Pour les personnes dans les ménages composés exclusivement de personnes longitudinales et pour les personnes longitudinales dans les ménages où tous les cohabitants sont IA} \\ w_{2003,p} \cdot \left(\dfrac{n_{L,h}}{n_{L,h} + n_{IP,h}} \right) \cdot & \text{Pour les personnes longitudinales dans les ménages composés de personnes longitudinales et d'au moins un cohabitant IP} \\ \dfrac{\sum_{h} w_{2003,p}}{n_{L,h} + n_{IP,h}} \cdot & \text{Pour tous les cohabitants} \end{cases}$$

où

$w_{2003,part}$ = Poids transversal partagé.

$w_{2003,p}$ = Poids transversal ajusté pour la combinaison des deux panels.

IP = Initialement présent.

IA = Initialement absent.

$n_{L,h}$ = Nombre de personnes longitudinales dans le ménage *h*.

$n_{IP,h}$ = Nombre de cohabitants initialement présents dans le ménage *h*.

Pour l'année de référence 2003, 75 206 individus ont un poids transversal intégré non nul et 70 278 individus ont un poids transversal travail non nul après l'étape du partage des poids.

5.6 Ajustement pour les valeurs influentes

Comme toute enquête essayant d'estimer la distribution des revenus des gens, l'EDTR est vraisemblablement confrontée au problème des observations influentes (ces distributions ayant de très longues queues). Les observations influentes ont un impact important sur les estimations provinciales du total et de la moyenne du revenu ainsi que sur leur estimation de variance. L'impact devient encore plus accentué sur les estimations de plus petits domaines comme les estimations selon le groupe d'âge, le sexe ou la taille des familles. Pour diminuer l'effet de ces observations sur les estimations, il faut d'abord trouver un moyen d'identifier ces observations, puis développer une méthode pour ajuster ces observations de telle sorte qu'elle permette de réduire l'influence de celles-ci sur les estimations.

Il est à noter que les étapes d'identification des observations influentes et de calcul des facteurs d'ajustement sont faites sur les poids intégrés seulement, mais que les facteurs d'ajustement dérivés sont appliqués aussi bien sur les poids intégrés que sur les poids travail. Avant la dernière révision historique, cette étape était faite en utilisant le poids individu (Lévesque et Franklin, 2000)

Une méthode de détection et de traitement des valeurs influentes a été mise sur pied par l'EFC (Tremblay, 1998). Comme on savait que l'EFC serait intégrée à l'EDTR à la collecte de 1999, la méthodologie développée a tenu compte du fait que l'enquête qui remplacerait l'EFC serait de nature longitudinale. L'approche de l'EFC a ainsi été adoptée par l'EDTR pour traiter le problème des revenus influents. La méthode consiste d'abord à identifier les observations influentes, c'est-à-dire celles qui contribuent trop fortement à l'estimation pondérée du total des revenus individuels de leur province. Il s'agit en général d'observations qui se trouvent dans une tranche de revenu élevé. Le revenu d'un individu (on parle ici du revenu total moins les gains en capital) est considéré comme influent si la contribution de son revenu pondéré par rapport à l'estimation pondérée provinciale dépasse un certain seuil. Ce seuil correspond à la borne inférieure de l'intervalle calculée à partir de la méthode des quartiles. Il se définit comme étant MED + k (Q3 - MED), où MED et Q3 sont respectivement la médiane et le troisième quartile de la contribution pondérée des

individus, et k est un paramètre fixe pour chaque province qui a été déterminé après avoir examiné les observations ayant les pourcentages de contribution les plus élevés à l'estimation provinciale. La valeur de k est de 25 pour toutes les provinces, exceptés pour le Québec et l'Ontario où k est égal à 75.

Une fois que les observations influentes sont identifiées, il faut maintenant déterminer les facteurs d'ajustement applicables aux poids associés à ces observations. Pour ce faire, on évalue si l'échantillon sur-représente les revenus supérieurs à 99 999$ par rapport à la distribution des revenus de la population des déclarants de données fiscales (distribution obtenue du fichier T1 de l'Agence du revenu du Canada). Si c'est le cas, les poids sont ajustés de sorte que la distribution de l'échantillon pondéré coïncide avec celle des données fiscales sauf dans le cas où cette différence est faible (ce qui vise à maintenir au minimum les ajustements de poids). L'ajustement des poids se fait en calculant un certain facteur d'ajustement compris entre 0 et 1. Plus précisément, voici comment le calcul du facteur d'ajustement est effectué pour chaque individu ayant un revenu influent. Notons que ces calculs sont faits au niveau provincial.

> "*Si une province contient plus d'un individu influent, le facteur d'ajustement est d'abord calculé pour l'individu influent avec le revenu (non pondéré) le plus élevé. Supposons que cet individu est dans la classe de revenu c. L'ajustement fera en sorte que l'estimation du nombre d'individus qui se trouvent dans la classe c ou dans toute autre classe de revenu supérieure, soit égale au nombre d'individus dans ces mêmes classes sur le fichier de données fiscales, sujet toutefois à certaines contraintes. D'abord, uniquement les poids des individus influents devront être modifiés dans cet ajustement. De plus, si l'ajustement entraîne une hausse du poids, c'est-à-dire que le facteur d'ajustement est supérieur à 1, l'ajustement ne sera pas appliqué et le poids initial sera conservé. Si le fichier de données fiscales indique qu'il n'y a aucun individu dans les classes de revenu, l'ajustement ramènera le poids à 1. De la même façon, on ramènera le poids à 1 si le nombre d'individus estimés dans ces classes à partir des individus non influents seulement, dépasse déjà le nombre d'individus selon le fichier de données fiscales.*" (Tremblay, 1998).

Le calcul du facteur d'ajustement est dépendant d'un individu à l'autre. Lorsque le facteur d'ajustement a été calculé pour l'individu influent avec le plus haut revenu, c'est le poids ajusté de cet individu qui sera considéré dans le calcul du facteur d'ajustement de l'individu suivant.

Bien que le facteur d'ajustement soit calculé au niveau de l'individu, il devra être appliqué au poids de la personne identifiée comme influente ainsi qu'aux poids de toutes les autres personnes habitant le même ménage. Pour les ménages où aucun individu n'a été traité pour la valeur influente de son revenu, le facteur d'ajustement est de 1 pour tous les membres de ces ménages.

Afin de détecter les valeurs influentes, la méthode utilise le poids obtenu après l'étape du calage (dont il est question à la section suivante) puisque cette dernière repose sur la distribution des revenus de la population. Il faut donc faire un calage temporaire avant d'effectuer l'étape de l'ajustement pour les valeurs influentes, puis utiliser les poids temporaires calés pour identifier et traiter les observations influentes. Une fois les facteurs d'ajustement calculés, l'étape du calage doit être refaite en appliquant ces facteurs aux poids que l'on avait avant le calage temporaire (donc aux poids transversaux obtenus à la suite du partage des poids, $w_{2003,\,part}$).

Le poids transversal après ajustement pour les valeurs influentes est calculé comme suit :

$$w_{2003,infl} = w_{2003,\,part} * b_{infl}$$

où $w_{2003,infl}$ = Poids transversal ajusté pour les valeurs influentes

$w_{2003,\,part}$ = Poids transversal obtenu à la suite du partage des poids

b_{infl} = Facteur d'ajustement pour les valeurs influentes.

En incluant les individus qui habitent avec toutes les personnes identifiées comme ayant des revenus influents à la vague courante ou aux vagues précédentes, l'ajustement pour les valeurs influentes a été appliqué à un total de 66 personnes pour l'année de référence 2003. Les facteurs d'ajustement apportés aux poids de ces personnes varient de 0,003 à 0,98.

5.7 Calage sur marges

La dernière étape de la pondération transversale est l'étape du calage sur marges. Tel que mentionné dans la section 4.5, cette étape de la pondération a subit de nombreux changements depuis les débuts de l'EDTR. Ceux-ci ont principalement été apportés lors des deux dernières révisions historiques qui ont eu lieu en 2002 et en 2005 suite à un vaste projet d'harmonisation du calage pour les enquêtes sur les dépenses et les revenus.

Rappelons que le but du calage est de faire en sorte que les estimations d'enquête pour certains sous-groupes de l'échantillon correspondent à des totaux connus de la population pour une année donnée. Or, à partir de l'année 1993, des écarts grandissants entre certaines estimations de l'enquête et celles provenant de sources externes ont été observés. Ces écarts, ainsi que certaines différences entre les estimations des diverses enquêtes sur le revenu et les dépenses ont été à l'origine du lancement en 1999 du Projet de calage harmonisé de la statistique du revenu. L'objectif de ce projet était d'harmoniser les stratégies de calage des enquêtes sur le revenu et les dépenses afin d'améliorer la comparabilité entre les diverses estimations tirées des enquêtes et celles provenant de sources externes (Tremblay, 2005).

Pour l'EDTR, il y avait deux problèmes importants à résoudre en ce qui concerne les estimations d'enquête. Le premier touchait la répartition de l'échantillon selon la taille des ménages et des entités économiques. En comparant les estimations tirées de l'EDTR avant calage sur les totaux connus de la population (appelé glissement), on observait une sous-estimation des ménages de taille deux et plus supérieure à 8 % alors que le nombre de ménages de taille 1 n'était sous-estimé que d'environ 2% (LaRoche, 2005). Pour ce qui est des entités économiques, on remarquait aussi une sous-estimation du nombre d'entités économiques pour les différentes tailles, la plus importante étant de loin celle des entités économiques de taille 1 qui s'élevait à près de 17%. Ce déséquilibre entre les estimations du nombre de ménages et d'entités économiques selon la taille résultait en une sous-estimation très importante du nombre d'entités économiques de taille 1 vivant dans les ménages de taille deux et plus. Lorsqu'aucun contrôle selon la taille de ménage et d'entité économique n'était utilisé dans le calage, on observait alors un biais important dans les estimations correspondantes de l'enquête.

Le deuxième problème observé en terme de représentativité de l'échantillon concernait le revenu. À partir des années 1990, on observait des écarts grandissants entre les estimations sur le revenu tirées des enquêtes et celles provenant de sources externes fiables. En effet, lorsqu'on comparait la série sur les salaires et traitements agrégés tirées de l'EFC/EDTR à celle provenant du Système de comptabilité nationale, on pouvait constater qu'entre 1980 et 1991, l'EFC surestimait les salaires agrégés d'environ 1.6% à 3.6% par rapport aux estimations du Système de comptabilité nationale et qu'à partir de 1992, cet écart s'agrandissait pour fluctuer entre 3.1% et 6.2% (Lathe, 2005).

La répartition des salariés selon les classes de salaires et traitements a donc été examinée de façon à identifier plus précisément l'origine de cette surestimation des totaux de salaires et traitements agrégés. Selon les résultats obtenus, la tendance de l'EDTR était de surestimer le nombre de personnes dans les fourchettes du milieu de la répartition des salaires et traitement et de sous-estimer le nombre de personnes dont les salaires et traitements se situaient aux deux extrémités de la répartition. Cette tendance était aussi présente pour l'EFC et pour d'autres enquêtes (Tremblay, 2005).

Dans le cadre du Projet sur le calage harmonisé de la statistique du revenu, des nombreuses études ont été effectuées de façon à identifier la meilleure stratégie de calage, c'est-à-dire celle qui aurait pour effet de réduire les écarts entre certaines estimations-clés provenant de différentes sources tout en limitant l'ampleur de la distorsion dans les poids. Un autre critère important dans le choix de la meilleure stratégie était que

celle-ci se devait de minimiser les impacts négatifs sur d'autres estimations d'enquête non contrôlées par le calage.

Un grand nombre de totaux de contrôle ont été évalués de façon à sélectionner ceux étant les plus appropriés pour l'enquête. En ce qui a trait aux totaux démographiques, des totaux de contrôle tels que le nombre de personnes selon divers groupes d'âge et de sexe, le nombre de ménages et d'entités économiques selon la taille et selon le type ont été évalués. Pour les totaux de contrôle basés sur le revenu, différentes sources externes ont été considérées, comme par exemple les fichiers T1 et T4 provenant de l'Agence du revenu du Canada. Le fichier T1 contient les données figurant dans la déclaration de revenus et de prestations que produisent les particuliers, alors que le fichier T4 contient les données figurant dans l'état de rémunération payée produit par les employeurs. Les nombreuses études réalisées ont démontré que le fichier T4 offrait une meilleure couverture, surtout en ce qui concerne les personnes à faibles revenus (Auger et Tremblay, 2005). Les totaux de contrôle retenus ont donc été ceux du nombre de salariés par classe de salaires et traitements. Les limites des classes ont été définies à partir de centiles spécifiques. Pour l'EDTR, les centiles utilisés sont le 10^{e}, le 25^{e}, le 50^{e}, le 65^{e}, le 75^{e} et enfin le 95^{e}, 98^{e} ou 99^{e}, selon le nombre d'observations incluses dans la dernière classe. Il est à noter que la limite de classe définie par le 10^{e} centile a été ajoutée pour l'EDTR en raison de l'importance de la population des faibles revenus pour l'enquête. Le choix des autres centiles a été surtout guidé par le fait qu'une bonne représentativité des hauts revenus était importante pour une meilleure estimation des totaux agrégés.

Pour l'EDTR, les améliorations au calage ont été implantées en deux phases. La première phase du projet a été introduite lors de la diffusion de l'année de référence 2000, soit en mai 2002. Les changements apportés à la stratégie de calage ne touchaient que les totaux démographiques. La deuxième phase du projet d'harmonisation du calage a été implantée lors de la diffusion des données de l'EDTR pour l'année de référence 2003, soit en mai 2005. La plus importante modification apportée au calage lors de cette révision est celle de l'ajout des totaux de contrôle basés sur le nombre de salariés par classe de salaire. Le nombre de groupes d'âge a aussi été réduit afin de diminuer le nombre total de contrôles utilisés et ainsi limiter la distorsion des poids lors du calage. Les totaux de contrôle ayant été utilisés lors de chacune des révisions historiques sont présentés dans le tableau 4.3 de la section sur la pondération longitudinale. Pour plus de détails sur les motifs ayant menés au choix de la stratégie de calage de l'EDTR actuelle et sur l'utilisation de données administratives à l'étape du calage, se référer à LaRoche (2005) et Latouche (2005).

Dans le cas du calage transversal, les estimations d'enquête doivent être ajustées de façon à représenter la population cible au 31 décembre de l'année de référence. Le classement des individus (éligibles transversalement et répondants) dans les différentes classes de calage se fait donc en utilisant leurs caractéristiques au 31 décembre de l'année de référence. Alors que le calage longitudinal doit être effectué en deux étapes, le calage transversal ne requiert qu'une seule étape puisque les totaux de contrôle ainsi que les variables d'enquête sont disponibles pour la période de référence. Les totaux de contrôle utilisés pour le calage transversal sont présentés dans le tableau 5.2.

Pour l'EDTR, les restrictions additionnelles suivantes sont appliquées aux poids finaux lors du calage. D'abord, les poids finaux doivent être supérieurs ou égaux à un pour les poids intégré et travail. De plus, pour le poids intégré, tous les membres d'un ménage doivent avoir des poids égaux suite au calage (Lemaître et Dufour, 1987). Rappelons que l'intégration des poids au niveau du ménage permet entre autres de produire des estimations cohérentes au niveau des ménages, des entités économiques et des personnes.

Tableau 5.2 Totaux de contrôle utilisés pour le calage transversal pour l'année de référence 2003

Totaux de contrôle			
Période de référence: 31 décembre 2003	Groupe d'âge selon le sexe	0 à 6 (les deux sexes) 7 à 17 18 à 24 25 à 34	35 à 44 45 à 54 55 à 64 65 et plus
	Effectifs selon la taille du ménage	Taille=1 et taille=2	
	Effectifs selon la taille de l'entité économique	Taille=1 et taille=2	
	Classes de salaires et traitements	7 classes définies selon les 10^e, 25^e, 50^e, 65^e, 75^e, 95^e ou 98^e ou 99^e centiles	

En utilisant les totaux de contrôle présentés dans le tableau 5.2 et en suivant la théorie présentée à la section 4.5.1, le vecteur des poids transversaux intégrés obtenu à la suite du calage est donné par

$$\underline{w}_{cal.int} = \text{diag}\left(\underline{w}_{2003.infl}\right)\underline{g}$$

où $\text{diag}\left(\underline{w}_{2003.infl}\right)$ est une matrice diagonale de dimension *n* x *n* avec les poids intégrés résultant de l'ajustement pour les valeurs influentes, $\underline{w}_{2003.infl}$, sur la diagonale, et $\underline{g}$ représente le vecteur des *g*-poids de dimension *n* x *1* obtenus en utilisant la contrainte additionnelle faisant en sorte que les poids finaux soient égaux pour tous les membre d'un ménage.

Pour le poids travail, le même processus est utilisé mais sans la contrainte des poids égaux à l'intérieur d'un même ménage. Le vecteur des poids transversaux travail obtenu à la suite du calage est donné par

$$\underline{w}_{cal.lab} = \text{diag}\left(\underline{w}_{2003.infl}\right)\underline{g}$$

où $\text{diag}\left(\underline{w}_{2003.infl}\right)$ est une matrice diagonale de dimension *n* x *n* avec les poids travail résultant de l'ajustement pour les valeurs influentes, $\underline{w}_{2003.infl}$, sur la diagonale, et $\underline{g}$ représente le vecteur des *g*-poids de dimension *n* x *1* résultant du calage.

Pour l'année de référence 2003, les poids transversaux intégrés après l'ajustement pour les valeurs influentes ont été en moyenne multipliés par 1,16 pour permettre le redressement de l'échantillon de l'EDTR. Pour la pondération travail, un ajustement moyen de 1,12 a été apporté aux poids transversaux travail pour effectuer le redressement de l'échantillon.

Les deux ensembles de poids résultant du calage sur marges constituent les ensembles des poids finaux transversaux de l'EDTR[10]. Le tableau 5.3 donne les médianes des poids transversaux finaux intégrés et travail, pour l'année de référence 2003. Seuls les individus avec des poids finaux non nuls sont utilisés dans le calcul de ces médianes. La province donnée dans le tableau fait référence à la province de résidence au 31 décembre 2003.

10. Auparavant, l'EDTR produisait aussi des ensembles de poids transversaux externes auxquels on ajoutait du bruit de façon aléatoire, et ce pour usage externe (Lévesque et Franklin, 2000). Ceux-ci ne sont plus produits.

Tableau 5.3 Médianes des poids transversaux finaux pour l'année de référence 2003

Province	Médiane	
	intégré	travail interne
	poids	
Terre Neuve et Labrador	154	160
Île-du-Prince-Édouard	65	68
Nouvelle-Écosse	162	174
Nouveau-Brunswick	163	176
Québec	342	373
Ontario	377	407
Manitoba	189	213
Saskatchewan	162	176
Alberta	453	493
Colombie-Britannique	561	598

6 Changements à venir et développements futurs

Afin d'améliorer la qualité et la fiabilité des estimations, les étapes de pondérations longitudinale et transversale de l'EDTR sont régulièrement révisées. Ces révisions sont effectuées à la suite d'études ou d'évaluations, en tenant compte de toute modification ayant trait aux objectifs, au traitement et à la diffusion. Des évaluations touchant diverses étapes de la pondération sont continuellement en cours afin d'être au fait des nouvelles méthodologies et d'utiliser les méthodes les plus appropriées pour les besoins de l'enquête.

L'EDTR, ainsi que plusieurs autres enquêtes, fait face à des taux de réponse de plus en plus bas et ce problème tend à s'aggraver. De nouvelles procédures ont été mises en place lors de la collecte de janvier 2006 de l'EDTR afin de cibler plus adéquatement les différents types de répondants pour tenter d'augmenter la probabilité de les rejoindre. Par exemple, si un type de ménage est plus susceptible d'être contacté à une période précise de la journée alors les contacts pour ce type de ménage sont priorisés durant cette période.

Même si on tente de réduire les effets de la non-réponse en améliorant les méthodes de collecte, il demeure important de s'assurer que les méthodes d'ajustement de non-réponse employées sont pertinentes et adéquates. Entre autres, différentes études doivent être effectuées prochainement afin d'évaluer la stratégie actuelle d'ajustement de la non-réponse. On examinera plus particulièrement l'ajustement de la non-réponse en fonction des aspects suivants : l'utilisation de la modélisation par segmentation par rapport à d'autres méthodes telles que la régression logistique; un ajustement fait au niveau provincial ou régional au lieu de national; l'ajout de variables explicatives selon le type de non-réponse (refus, non contact, etc.); l'ajout d'autres variables indicatrices basées sur la géographie pour expliquer la non-réponse. Ces études permettront d'établir s'il y a lieu d'apporter des changements à la méthodologie actuelle et d'implanter des nouvelles méthodes.

Un autre changement qu'il serait souhaitable d'apporter à l'enquête est le développement et la mise en œuvre d'une stratégie d'imputation des données sur le travail. Ainsi, la non-réponse à l'entrevue sur le travail serait compensée en imputant les données sur le travail et il ne serait donc plus utile de produire un poids travail. Comme il est plus complexe de développer une stratégie d'imputation des données sur le travail que sur le revenu, et qu'un tel projet requiert d'importantes ressources tant du côté méthodologique que sujet-matière, ce projet a été délaissé depuis de nombreuses années mais demeure toujours important pour le groupe de méthodologie. Tant et aussi longtemps que cette stratégie d'imputation des variables reliées au travail ne sera implantée, le poids travail restera nécessaire et continuera à être produit. Cependant, des changements pourraient être apportés à la définition employée pour définir un répondant dans le contexte du poids travail afin de limiter le nombre de valeurs manquantes pour les utilisateurs. Différentes options sont présentement à l'étude. Entre autres, on évalue la possibilité de changer la définition de répondant au niveau individuel au lieu de ménage. De plus, dans l'éventualité d'une nouvelle diffusion plus hâtive d'une publication portant sur les données reliées au travail, la nécessité de produire un poids travail plus rapidement doit être étudiée.

L'EDTR existe depuis maintenant plus de 10 ans. Suite aux résultats obtenus, aux expériences acquises, à la rétroaction des principaux utilisateurs ainsi qu'aux nouveaux besoins de données, un remaniement en profondeur de l'enquête est à l'étude et devrait être mise en œuvre dans les prochaines années. Parmi les caractéristiques de l'enquête qui seront réévaluées, on notera, la durée des panels, le nombre d'années avant l'introduction d'un nouveau panel ainsi que le nombre de panels se chevauchant.

7 Conclusion

L'objectif de l'EDTR de fournir des données longitudinales et transversales pour plusieurs caractéristiques reliées à la dynamique du travail et du revenu, et qui soient représentatives de la population des dix provinces canadiennes, est un grand défi. Les pondérations longitudinale et transversale expliquées en détails dans ce document permettent de produire des poids longitudinaux et transversaux nécessaires à ces fins. Pour produire ces poids, plusieurs étapes sont requises : la détermination des poids initiaux, l'ajustement pour la non-réponse, les ajustements pour la migration interprovinciale (seulement pour les poids transversaux), la combinaison des panels et le partage des poids (seulement pour les poids transversaux), l'ajustement pour les valeurs influentes et finalement le calage sur marges. Au fil des ans, ces étapes ont été améliorées et optimisées mais avec l'évolution des stratégies de traitement et de diffusion, de nouvelles mises au point sont nécessaires pour, entre autres, préserver la qualité et la fiabilité des données.

Remerciements

L'auteure tient à remercier Caroline Cauchon, Michel Latouche, Pierre Lavallée et Soumaya Moussa pour leurs commentaires et précieux conseils qui ont permis d'améliorer la qualité de ce document.

Annexe A : Liste des variables utilisées dans le modèle d'ajustement pour la non-réponse

Les variables dichotomiques suivantes ont été utilisées pour modéliser la non-réponse pour la cinquième vague du panel 3 et la deuxième vague du panel 4.

Educ1	- 0 à 8 années de scolarité
Educ2	- 9 à 13 années de scolarité (études secondaires)
Educ3	- études post-secondaires excluant le collège communautaire, le CÉGEP et l'université
Educ4	- collège communautaire, CÉGEP ou université (incluant doctorat)
RENTER	- 1=propriétaire, 0=locataire
EMPLOYE	- personnes occupées
NON_EMPL	- chômeurs
NLF	- inactifs
MARIE	- marié ou vivant en union libre
SEP_DIV	- séparé ou divorcé
VEUF	- veuf ou veuve
CELIB	- célibataire, n'a jamais été marié
P10 – P59	- une variable par province
SEXE	- sexe
AGE_0	- âge < 16
AGE_1	- age >=16 et <19
AGE_2	- age >=19 et <25
AGE_3	- age >=25 et <35
AGE_4	- age >=35 et <45
AGE_5	- age >=45 et <55
AGE_6	- age >=55 et <65
AGE_7	- age >=65
HHSZ1-HHSZ5	- nombre de personnes dans le ménage (HHSZ5=>5)
FAMTYPE1	- personne non apparentée
FAMTYPE2	- famille époux-épouse ou famille monoparentale sans jeune enfant
FAMTYPE3	- famille époux-épouse ou famille monoparentale avec des enfants âgés de 0 à 17 ans
FAMTYPE4	- autre genre de famille
STUDENTS	- 1=étudiant (à temps plein ou à temps partiel), 0=n'est pas un étudiant
CWORK1	- catégorie de travailleur = employé du secteur public
CWORK2	- catégorie de travailleur = employé du secteur privé
CWORK3	- est un travailleur indépendant du secteur privé avec ou sans employés ou est un travailleur familial non rémunéré
MULTIHH	- ménage à familles multiples
URBAIN	- Habite dans une région urbaine (basé sur le secteur de dénombrement)

Bibliographie

Auger, S. et Tremblay, J. (2005). *Évaluation de la qualité des estimations démographiques et des données fiscales utilisées dans le calage des différentes enquêtes à Statistique Canada.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.

Bankier, M. (2002). *Pondération pour le recensement du Canada 2001.* Modélisation de données d'enquête pour la recherche économique et sociale, Assemblée annuelle de la Société Statistique du Canada, juin 2002

Cotton, C., Bishop, K., Giles, P., Hewer, P. et Saint-Pierre, Y. (1999). *Comparaison des résultats de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de l'Enquête sur les finances des consommateurs (EFC) 1993-1997 : mise à jour*, publication de Statistique Canada, 75F0002MIF – No 99007.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindeyer, J. (1998). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526-XPB.

Godbout, S. (2004). *Étude sur l'ajustement pour la non-réponse et la surestimation des ménages individuels*, document interne, Statistique Canada.

Kalton, G. et Kasprzyk, D. (1986). *Le traitement des données d'enquête manquantes*, Techniques d'enquête, vol. 12, no 1, pp. 1-17.

LaRoche, S. (2005). *Stratégie de calage de l'Enquête sur la dynamique du travail et du revenu.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.

Lathe, H. (2005). *Enquête sur la dynamique du travail et du revenu : révision historique de 2003*, Série de document de recherche sur le Revenu, Statistique Canada, 75F0002MIF au catalogue - No 009

Latouche, M. (2005). *Leçons à tirer de l'utilisation de données administratives au stade de la pondération.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.

Latouche, M., Dufour, J. et Merkouris, T. (2000). *SLID Cross-Sectional Weighting: Combining Two or More Panels*, Série de document de recherche sur le revenu, Statistique Canada, 75F0002MIE – 00006.

Latouche, M. et Michaud, S. (1997). *Concerns pertaining to weighting of longitudinal surveys*, Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.

Lavallée, P. (1994). *Ajout du second panel à l'EDTR: sélection et pondération*, document interne, Statistique Canada.

Lavallée, P. (1995). *Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage de poids*, Techniques d'enquête, vol. 21, no 1, pp. 27-35.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée de partage des poids*, Édition de l'université de Bruxelles.

Lavigne, M. et Michaud, S. (1998). *Aspects généraux de l'enquête sur la dynamique du travail et du revenu*, document de travail sur la dynamique du revenu et du travail, Statistique Canada, 75F0002M No. 98-05.

Lévesque, I. et Franklin, S. (2000). *Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu, année de référence 1997*, document de travail sur la dynamique du revenu et du travail, Statistique Canada, N° 75F0002MPF-00004.

Lemaître, G. et Dufour, J. (1987). *Une méthode intégrée de pondération des personnes et des familles*, Techniques d'enquête, vol. 13, no 2, pp. 211-220.

Merkouris, T. (1999). *Cross-sectional Estimation in Multiple-Panel Household Surveys*, document de travail de la direction de méthodologie, Statistique Canada, HSMD-99-004E.

Naud, J,-F. (2004). *Pondération longitudinale avec panels combinés, 'Enquête sur la dynamique du travail et du revenu, 1996-2002.* Document de travail sur la dynamique du revenu et du travail, Statistique Canada N° 75F0002MIF2004008

Naud, J,-F. (2004). *Étude sur des modifications possibles à l'ajustement pour la non-réponse*,document interne, Statistique Canada.

Särndal, K. et coll (1992). *Model-Assisted Survey sampling*, New York: Springer-Verlag.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526.

Statistique Canada (1997). *Répartition du revenu au Canada selon la taille du revenu*, publication de Statistique Canada N 13-207-XPB, pp. 52-59.

Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. et Nadon, S. (1998). *Traitement de la non-réponse du cycle deux de l'enquête sur la santé de la population*, Techniques d'enquête, vol. 24, no 2, pp. 159-169.

Tremblay, J. (1998). *Détection des observations influentes pour l'enquête sur les finances des consommateurs (EFC) et l'enquête sur la dynamique du travail et du revenu (EDTR)*, document interne, Statistique Canada.

Tremblay, J. (2005). *Aperçu de la stratégie de calage harmonisée des statistiques du revenu de Statistique Canada.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.